The Impact of Paternal Metabolic Health on Sperm DNA Methylation and Fetal Growth

Thesis presented for the degree of Doctor of Philosophy in the Faculty of Population Health Sciences, University College London

Dr Karin Ingrid Fredrika Åsenius

Signed Declaration

I, Karin Ingrid Fredrika Åsenius, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgements

I owe my deepest gratitude to Dr David Williams for giving me the opportunity to undertake this project, and for tirelessly supporting me along the way. I feel extremely fortunate to have been able to conduct research in this fascinating and important field, and I hope that we can continue our collaboration for years to come.

I am also incredibly grateful to my subsidiary supervisor, Professor Vardhman Rakyan, for his constructive guidance and expertise.

I was fortunate to be awarded funding for the project from the Rosetrees Trust through a PhD stipend. This project would not have been achievable without their generous support.

From the bottom of my heart, I would like to thank Mia and Calle, who have lovingly supported me throughout all my endeavours. This is as much your achievement as it is mine. And I would like to thank Ludvig, who never fails to make me smile.

I would like to thank my friends, colleagues and fellow yogis for support, advice and much needed distraction. I would especially like to mention Dr Sarah Marzi and Dr Amy Danson, without whose encouragement and expertise the timely completion of this work would not have been possible.

I am forever grateful to Dr Anthony Swallow for his unwavering support which allowed me to embark on this rewarding academic journey.

Lastly, I would like to thank Tom, who has made these years happier than I could ever have hoped for. The best thing about the future is the privilege of spending it with you.

Abstract

Low birth weight is associated with cardiovascular disease and T2DM in later life. Paternal obesity and T2DM have been associated with an increased risk of fathering low birthweight offspring. Obesity is associated with epigenetic changes in blood, but few studies have replicated DNA methylation differences found in obese subjects. Animal studies have shown that obesity and insulin resistance are associated with DNA methylation changes in sperm, which in turn could mediate intergenerational effects. Such findings are lacking in humans. My PhD explored the association between paternal metabolic traits and the birth weight of his offspring. I then investigated whether DNA methylation signatures in spermatozoa of obese fathers could underlie any observed association with his offspring birthweight.

First, I performed a prospective cohort study of 500 mother-father-offspring trios to identify paternal metabolic traits associated with an increased risk of fathering low birth weight offspring. Out of 390 trios, including 64 obese men and 48 growth-restricted offspring, I did not discover any significant paternal metabolic traits associated with fathering low-birthweight offspring. However, I found that paternal (own) birth weight is associated with the birth weight of his offspring. This suggests that paternal genetic factors are more influential in determining his offspring's growth *in utero* than are factors acquired during his lifetime.

Second, I performed a systematic review of studies that had investigated DNA methylation in human sperm. From this review, I summarised current knowledge and generated recommendations for future research.

I then performed the largest characterisation of matched human sperm and blood samples to date using the most comprehensive DNA methylation profiling array, the MethylationEPIC array. Results showed that the DNA methylomes of sperm and blood are highly discordant and in effect completely uncorrelated. Future studies of intergenerational effects will have to study germ cells, rather than blood.

Lastly, I attempted to validate previously-identified DNA methylation signatures associated with male obesity. Despite comparing 96 well-characterised obese men with 96 lean men, I was unable to replicate any previously identified differentially methylated CpG sites associated with obesity, in their blood. In a linear regression model, I identified two CpG sites, cg07037944 and cg26651978, as being suggestive of an association with BMI. These results will contribute to a larger cohort study of 1000 obese and 1000 lean men that aims to identify a robust and reproducible DNA methylation profile associated with obesity.

In conclusion, this thesis did not prove my pre-determined hypotheses. However, it does present findings which advance our understanding of the intriguing possibility that acquired parental metabolic phenotype may influence offspring birthweight through intergenerational inheritance of epigenetic marks.

Impact Statement

Research presented in this thesis shows that a father's Body Mass Index (BMI) around the time of conception has little impact on the birth weight of his offspring. However, other research has suggested that paternal BMI instead has a greater influence on the BMI of his offspring during childhood and later life. Research following on from this thesis will follow up the children born during my study to determine whether periconceptual paternal obesity sets a template for offspring obesity. Public health policies that target men at risk of obesity around the time of his partner's pregnancy could have a beneficial impact on the health of the next generation and beyond.

I found that paternal (own) birth weight is associated with the birth weight of his offspring, but that there is no such association between maternal (own) birthweight and offspring birthweight. This observation raises the intriguing possibility that paternal influences on offspring growth are predominantly genetic, whilst maternal influences are predominantly environmental. Furthermore, customised birthweight centiles for estimating fetal weight in utero may become more accurate if they included paternal (own) birth weight. This possibility needs to be tested in a future project.

In the largest study of DNA methylation in matched sperm and blood samples performed to date, I show that the methylation profiles of sperm and blood are highly distinct and practically uncorrelated. These findings emphasise the necessity of studying germ cells, rather than proxy tissues, in analyses of intergenerational effects. They also question the validity of previous studies where blood has been used as a proxy tissue for sperm DNA methylation.

This thesis contains the first systematic review of studies of DNA methylation in human sperm. From this, I summarise current knowledge in the field and set out a series of recommendations for future research. The systematic review will be of value for anyone interested in the epigenome of human germ cells, or designing studies of

the human sperm methylome. Such studies should aid our understanding of fertility, embryogenesis and the potential for the spermatozoal epigenome to be influenced by acquired traits.

In terms of research methodologies that could be applied in future research, I have shown that bisulfite-PCR-sequencing represents a high-throughput, cost-effective, targeted alternative to reduced representation bisulfite sequencing in epigenome wide association studies. Results are well-correlated to those of DNA methylation profiling using the Illumina MethylationEPIC array. However, the methodology also has drawbacks, such as a generally lower read coverage.

Future prospective studies to understand drivers of fetal growth that might be shared by a father and his offspring, may provide an early opportunity for the primary prevention of obesity in the next generation.

Table of Contents

List of Figures

List of Tables

Abbreviations

A	Adenine
AGA	Appropriate for gestational age
ART	Assisted reproductive technologies
BMI	Body mass index
bp	Base-pair
bsDNA	Bisulfite converted DNA
C	Cytosine
CD	Control diet
CI	Confidence interval
COBRA	Combined bisulfite restriction analysis
CpG	Cytosine followed by guanine
dbSNP	Single nucleotide polymorphism database
ddNTP	Dideoxynucleotide triphosphate
DMR	Differentially methylated region
DNAm	DNA methylation
DNP	Dinitrophenyl
dNTP	Deoxyribonucleotide triphosphate
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
ESCs	Embryonic stem cells
EWAS	Epigenome wide association study
FGR	Fetal growth restriction
G	Guanine
GP	General practitioner
GWAS	Genome wide association study
HbA1c	Glycated haemoglobin
HDL	High density lipoprotein cholesterol
HFD	High fat diet
HM450	Infinium human methylation 450 array
HOMA	Homeostasis model assessment
HOMA-IR	Homeostatic model assessment of insulin resistance
IFC	Integrated fluidic circuit
IR	Insulin resistance
IUGR	Intrauterine growth restriction
IVF	In vitro fertilisation
LDL	Low density lipoprotein cholesterol
MODY	Maturity onset diabetes of the young
ND	Not detailed
OR	Odds ratio
P1	Protamine 1

P2	Protamine 2
PBL	Peripheral blood leukocyte
PCA	Principal component analysis
PGC	Primordial germ cell
PTM	Post-translational modification
QC	Quality control
RefSeq	NCBI reference sequence database
RNA	Ribonucleic acid
RPM	Revolutions per minute
SD	Standard deviation
SDS	Sodium dodecyl sulfate
SGA	Small for gestational age
sncRNA	Small non-coding RNA
SNP	Single nucleotide polymorphism
T	Thymine
T2DM	Type 2 diabetes mellitus
tRNA	Transfer RNA
TSS	Transcription start site
U	Uracil
UTR	Untranslated region
WB	Whole blood
WC	Waist circumference
WGBS	Whole-genome bisulfite sequencing

Chapter 1
Introduction

1.1 The Obesity Epidemic

The obesity epidemic and related metabolic disease constitute major public health problems globally. Obesity, defined as a Body Mass Index (BMI) exceeding 30 kg/m2, is a major risk factor for comorbities such as Type 2 Diabetes Mellitus (T2DM), cardiovascular disease, chronic kidney disease, musculoskeletal disorders and some cancers, and surpasses smoking and alcohol consumption in its negative effects on health (1-3). Worldwide in 2016, 39% of adults, more than 1.9 billion people, were overweight (BMI >25 kg/m2), one-third of whom were obese (2). Since 1975, the worldwide prevalence of obesity has nearly tripled, and the trend is still increasing. Indeed, a recent pooled analysis of trends in body mass index, which collated results from studies with over 19 million participants in total, suggested that unless current trends are halted, the global prevalence of obesity will reach 18% in men and surpass 21% in women by 2025 (4). In in low- and middle-income countries (LMICs), obesity increasingly occurs alongside persistent burdens of underweight, owing to changes in dietary habits and an increasingly sedentary lifestyle (5).

In 2016, more than one in four UK adults (26.2%) were obese, and a further 35.2% were overweight (6). As a consequence, the UK prevalence of Type 2 Diabetes (T2DM) has risen dramatically, estimated at almost 9% of the UK adult population (7). The condition is more prevalent in people of black and Asian ethnicities and those in lower socioeconomic positions (8). In addition to its detrimental effect on quality of life, the rise in metabolic disease is associated with a significant economic burden. Indeed, recent estimates of the cost for treating obesity and related morbidity in the NHS amount to nearly £6 billion annually, whereas the annual cost for treating Type 2 diabetes and its complications amounts to £14 billion pounds (3, 9). Together, healthcare spending for these two conditions account for approximately one sixth of the total NHS budget (10).

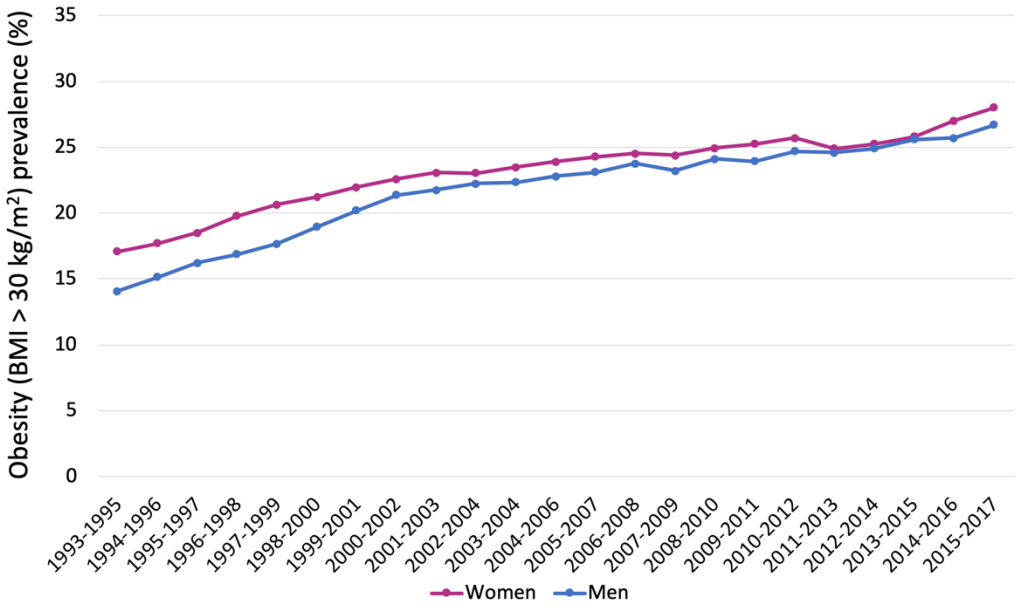


Figure 1-1. Trends in obesity prevalence among UK adults 1993-2017.
Data derived from annual Health Survey for England estimates of the prevalence of obesity among UK adult (over 16 years) men and women. Data points indicate three-year average values. Created using information available from (11).

1.1.1 Health Consequences of Obesity

Obesity and related disease constitute complex and heterogenous conditions resulting from an interplay of genetic, environmental and behavioural factors. Associated health risks are diverse, and include cardiovascular, pulmonary, endocrine, psychological and musculoskeletal conditions, as presented in Figure 1-2.

Among the conditions most closely associated with obesity is the development of insulin resistance (IR) and T2DM. Although the exact pathophysiology of IR is heterogeneous and still a topic of research, lifestyle factors such as an atherogenic diet and physical inactivity are significant risk factors for its development. Enduring hyperglycaemia causes a compensatory high secretion of insulin from pancreatic β cells, which leads to diminishing insulin sensitivity of liver, skeletal muscle and adipose tissue. Eventually, pancreatic β cells fail to secrete high enough levels of insulin to maintain adequate glucose levels, and a pre-clinical condition of insulin resistance becomes overt T2DM (12).

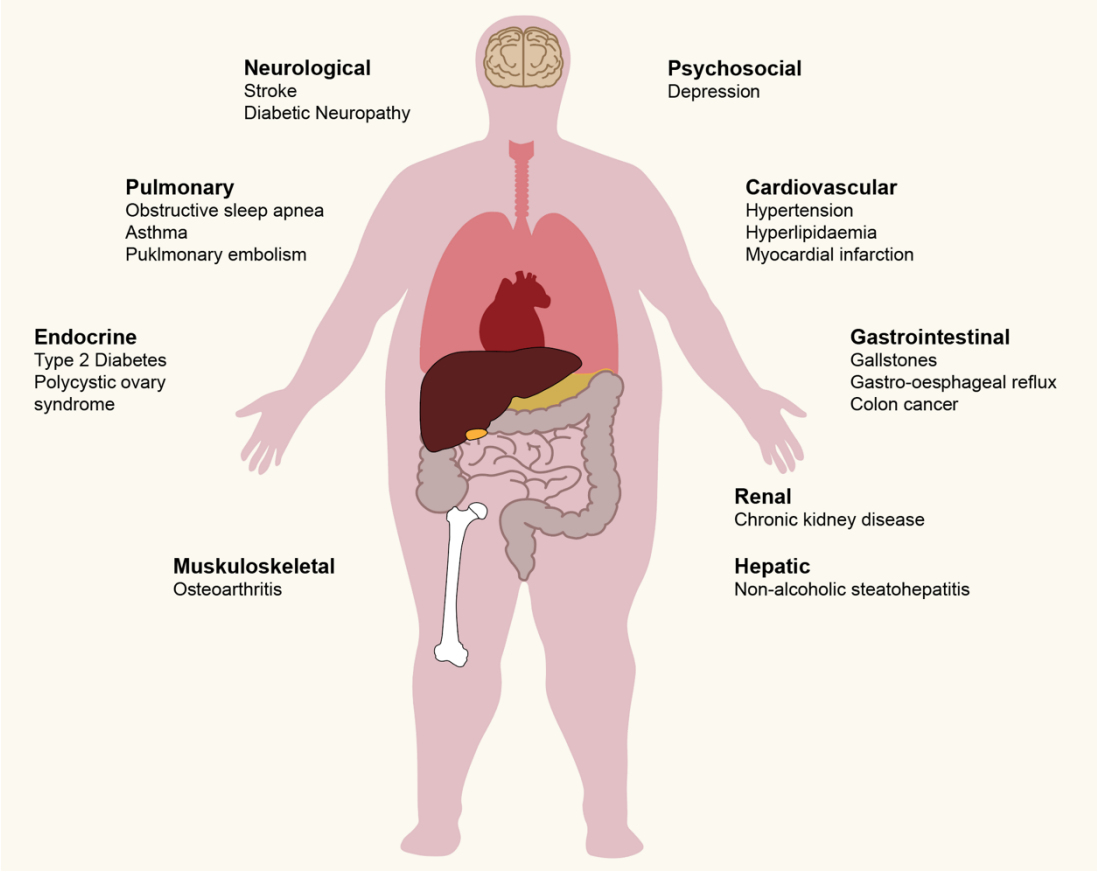


Figure 1-2. Selected health risks associated with obesity.
Figure created using information from (13, 14).

Obesity, in particular central adiposity, forms one of the core diagnostic criteria of the Metabolic Syndrome. This describes a cluster of conditions that together are associated with a threefold increased risk of a cardiovascular event such as a myocardial infarction (12) (Figure 1-3). According to International Diabetes Federation criteria, a diagnosis of metabolic syndrome requires central obesity (waist circumference ≥94 cm for Caucasian males) plus any two of the following: raised triglycerides (≥1.7 mmol/L or specific treatment for high triglycerides), reduced HDL cholesterol (<1.03 mmol/L in males or specific treatment for this lipid abnormality), raised blood pressure (systolic blood pressure ≥130 mmHg or diastolic blood pressure ≥80 mmHg or be treated for hypertension) or raised fasting plasma glucose (≥5.5 mmol/L or previously diagnosed T2DM) (12).

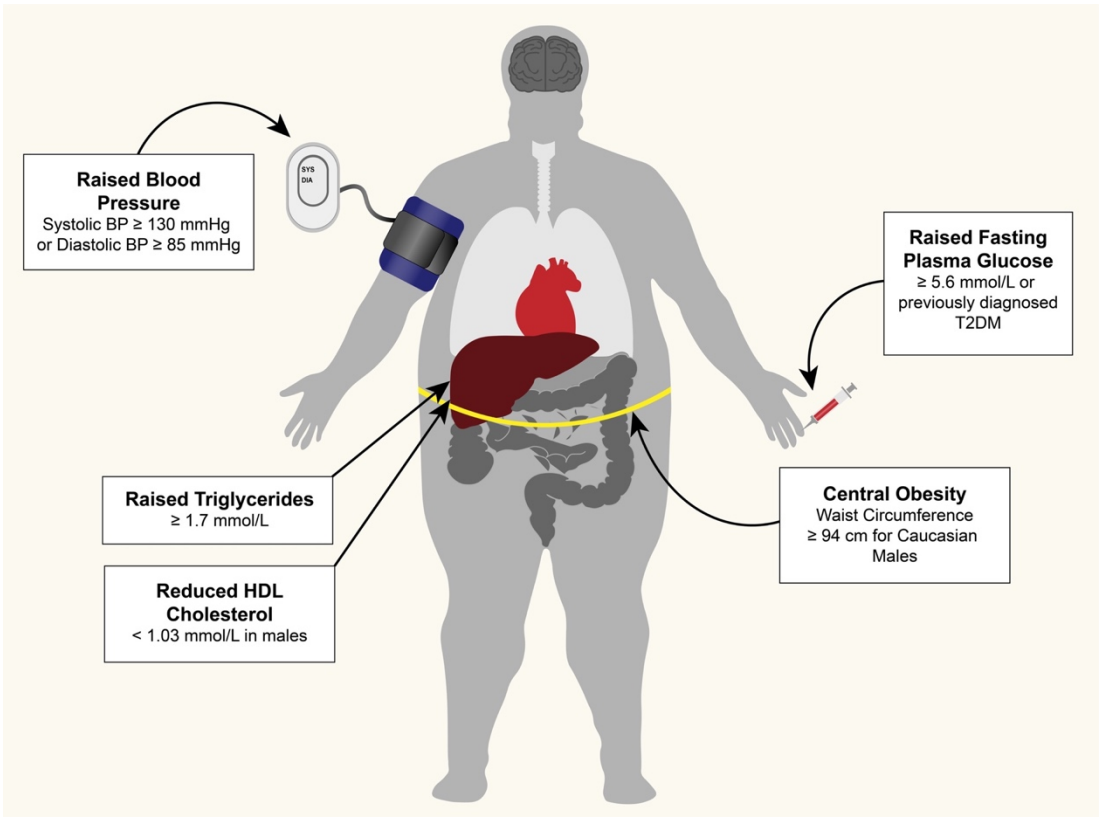


Figure 1-3. Diagnostic criteria for the Metabolic Syndrome according to the International Diabetes Federation (IDF) 2006.
Figure created using information from (12).

Obesity in pregnancy is associated with health risks both for the pregnant woman and the fetus. Thus, a pre-pregnancy BMI >40 kg/m^2 significantly increases the risk of gestational diabetes mellitus, gestational hypertension, pre-eclampsia and thromboembolic events during pregnancy. Further, maternal obesity is associated with an increased risk of adverse intrapartum events, including emergency caesarean sections, wound healing complications following caesarean sections and admissions to intensive care units. Fetal and neonatal risks associated with maternal obesity include an increased risk of fetal malformations, intrauterine death and macrosomia, the latter of which is discussed further in section 1.2 (15).

There is an estimated 4 million obesity associated deaths per year. Of these, 2.7 million are estimated to occur as a consequence of cardiovascular events, 0.6 million from T2DM-related events, and the rest from causes such as cancer, chronic kidney disease, musculoskeletal disorders and adverse pregnancy outcomes associated with maternal obesity (13).

1.1.2 Obesity and Metabolic Disease: The Importance of Prevention

Despite a nearly universal concern about the health risks associated with obesity and the enormous efforts that have gone into halting current trends in obesity prevalence, no national success stories have been reported (16). Indeed, once obesity is acquired, the probability of achieving and maintaining a normal body weight is low; an observational study using electronic primary care health records of over 150,000 obese individuals with a nine year follow-up period showed that the annual probability of attaining a normal weight was 1 in 210 for men and 1 in 124 for women. The study further showed that the probability of attaining a normal weight decreases with increasing BMI category, and that even among individuals who initially managed a decrease in BMI category, the vast majority thereafter regained weight (

Figure 1-4) (17).

Further, large-scale observational studies show that individuals born to obese parents have a significantly increased risk of themselves becoming obese as adults. Thus, in a pooled analysis of the annual Health Surveys for England carried out between 2001 and 2006, in total analysing data from 4,432 families, having two obese parents was associated with a significantly increased risk of childhood obesity (OR: 22.3; 95% CI: 10.3, 48.4; $p < 0.01$) independent of age, sex, socioeconomic status, and ethnicity (18). Childhood obesity is itself a strong predictor of adult obesity. Indeed, it has been estimated that when both parents are obese, about 80% of their children will be obese. This incidence falls to approximately 40% when one parent is obese, and to approximately 14% when both parents are lean (19).

It is therefore clear that in order to halt the obesity epidemic and its devastating consequences, focus needs to be placed on primary prevention. As stated by the Department of Health in a 2011 report on tackling obesity in the UK:

" We need to adopt a life course approach – from pre-conception, through pregnancy, infancy, early years, childhood, adolescence and teenage years, and through to adulthood and preparing for older age" (20).

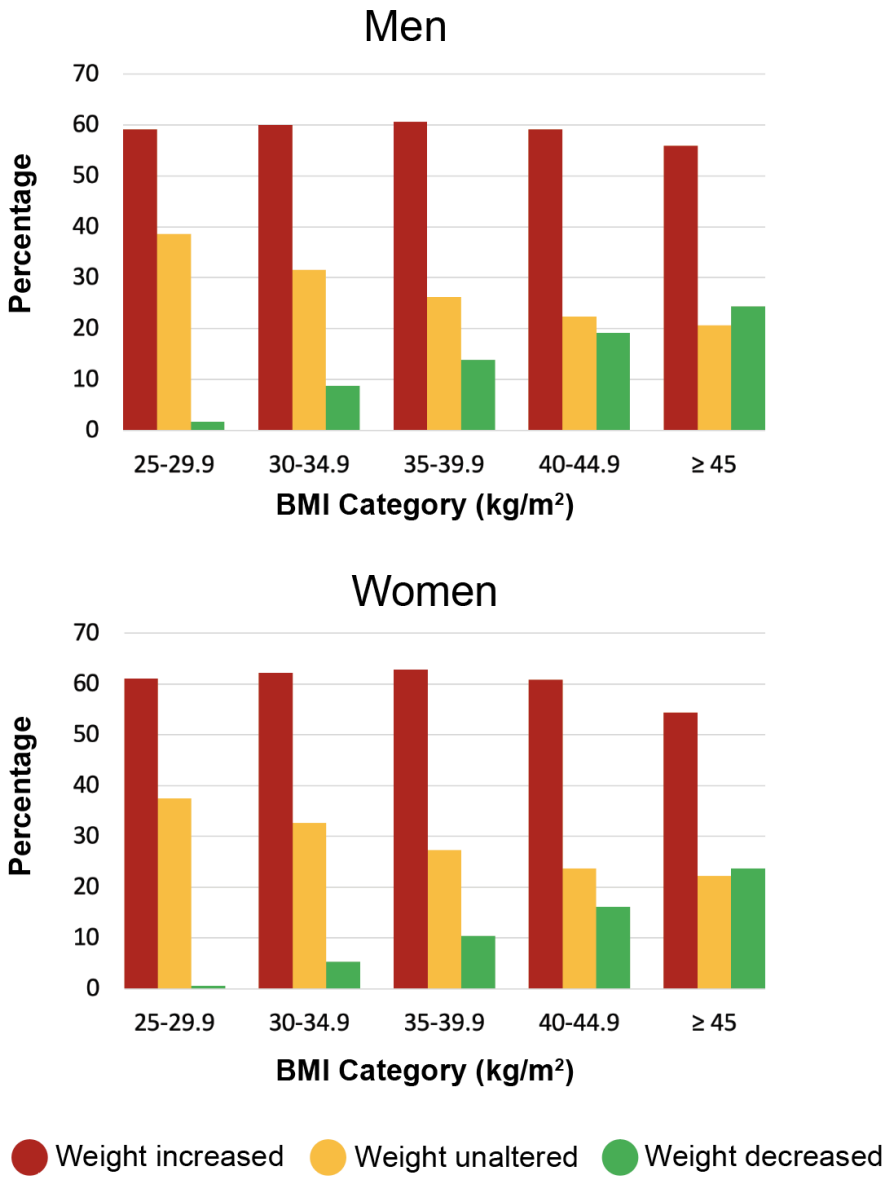


Figure 1-4. Probability of individuals who initially decreased in BMI category to thereafter increase, maintain or decrease their weight.
As visualised, approximately 60% of obese individuals who initially decreased their BMI category (to the category indicated on the x axis) thereafter increased their weight over the nine year follow up period. Data derived from (17).

1.2 The Genetic Architecture of Obesity and Type 2 Diabetes Mellitus

Obesity and T2DM are conditions characterised by high heritability estimates (21, 22). Heritability describes the proportion of the phenotypic variance of a trait that is attributable to genetic factors (23). Classically, this has been estimated by collecting data from twin, adoption and family studies, which all attempt to separate the genetic versus environmental influences on a trait based on genetic relatedness of the individuals included in the study (24, 25).

Studies in monozygotic twin pairs rely on the twins' identical genetic makeup, making any difference in phenotypic variance between a twin pair attributable to unique environmental factors. In contrast, dizygotic twin pairs only share approximately 50% of their variable genetic make-up and can be assumed to have shared the same environment (24). One can assume that unique environmental factors contribute to a trait equally in monozygotic and dizygotic twin pairs. Therefore, it is possible to estimate the effect of the additional genetic similarity between monozygotic twin pairs by comparing the phenotypic correlation between monozygotic and dizygotic twins respectively. In adoption studies, the genetic contribution to a trait is estimated by comparing the risk of the trait to biological versus adoptive relatives of affected versus control adoptees. If the adoption study is of monozygotic twins that have been separated at birth and reared apart, any phenotypic correlation between the twin pair can be assumed to arise from genetic factors (24). Family studies frequently use estimates of the expected genetic relatedness between two individuals (e.g. 50% between full siblings and 12.5% between first cousins) (25). In family studies of binary traits, the concordance or discordance of disease status can be estimated using a 2 by 2 contingency table. For continuous phenotypic traits, e.g. BMI, heritability can be estimated by the slope of the regression line, which approximates the heritability of a trait when the mean phenotypic value of the parents is used (24, 25). In summary, such pedigree analyses have allowed estimations of how much of the phenotypic variance in BMI and in the risk of T2DM can be attributed to genetic factors.

In a meta-analysis of studies involving over 34,000 monozygotic twin pairs in total, the Discordant Twin (DISCOTWIN) Consortium estimated the heritability of T2DM to

be 72% (95% CI 61-78%) (21). Similarly, in the case of obesity, heritability estimates range from 40% to 70% (22). The high prevalence, heritability and considerable health consequences of T2DM and obesity have spurred intensive efforts into uncovering genetic variants that increase the risk of developing these conditions.

The current understanding of the genetic basis of obesity and T2DM is similar to that of other common, complex diseases. Thus, in a small proportion of individuals, the condition occurs as a result of single gene variants with large effect sizes. However, in the vast majority of cases, the disease phenotype results from a complex interaction between lifestyle factors and a large number of disease risk conferring genetic variants, each with small effect sizes (26-28)(Figure 1-5).

At least 10 gene variants have been associated with severe, monogenic obesity (typically defined as a BMI > 35 kg/m^2) (29). These include single nucleotide polymorphisms (SNPs) in the gene for leptin (*LEP*), and in the Melanocortin 4 Receptor (*MC4R*). Leptin is a hormone primarily produced by adipocytes and is involved in appetite regulation in part by suppressing the orexigenic peptides neuropeptide Y (*NPY*) and agouti-related peptide (*AGRP*) in the hypothalamus, thus reducing food intake (30). The melanocortin 4 receptor is a hypothalamic receptor involved in a leptin-targeted neural circuit of energy homeostasis (29). Indeed, the melanocortin 4 receptor, which is present in the hypothalamic arcuate nucleus, responds to levels of leptin, glucagon-like peptide 1 (*GLP-1*) and ghrelin (amongst others) and has a pivotal role in integrating these signals and regulating energy expenditure, satiety and growth accordingly (31). Individuals with an *MC4R* mutation typically present with severe obesity in childhood, lack of feelings of satiety, and have a higher fat mass ratio as adults (32). Studies in mice have further demonstrated that a targeted disruption of the *MC4R* gene is associated with increased food intake, obesity and hyperinsulinaemia (33). However, the most common form of these genetic variants, which is mutations in *MC4R* causing melanocortin 4 receptor deficiency, is only present in approximately 1% of adults with a BMI exceeding 30 kg/m^2, and can thus only account for a minority of obesity cases (34).

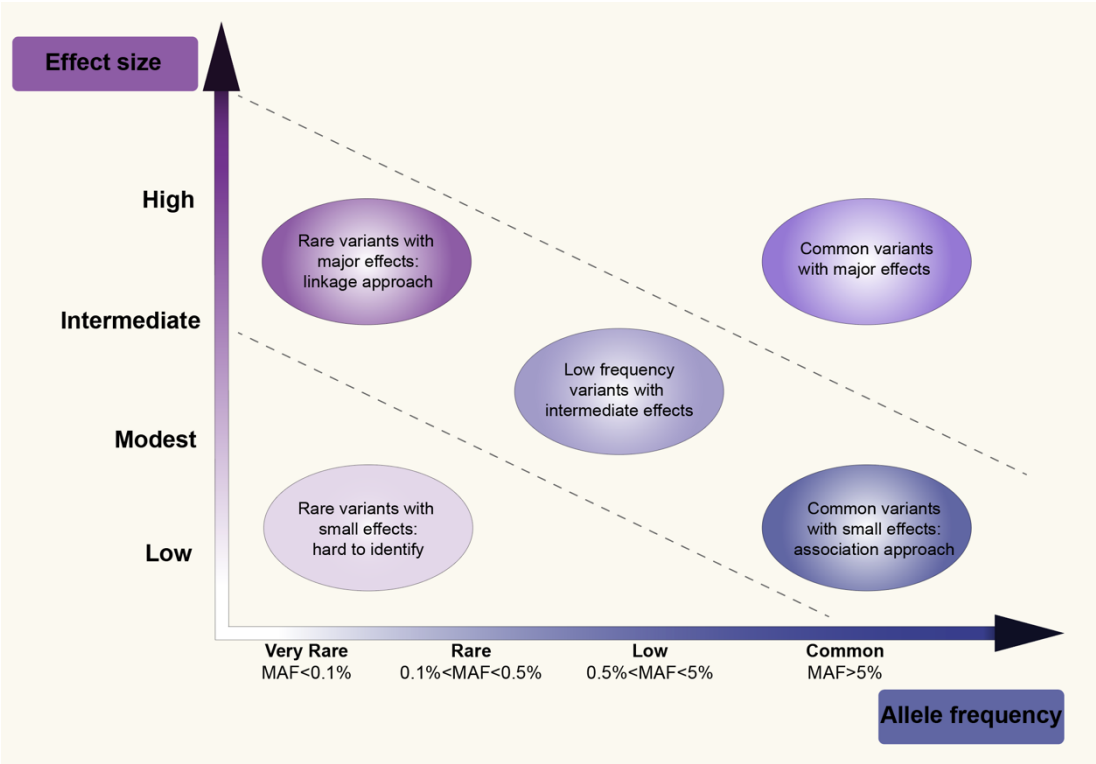


Figure 1-5. The genetic architecture of common conditions such as T2DM and obesity, and approaches for identifying risk conferring alleles.
The vast majority of cases of obesity and T2DM are likely to have arisen from a complex interaction between a large number of disease-conferring alleles, each with small effect sizes, that together with lifestyle factors produce the disease phenotype. MAF = Mean Allele Frequency. Figure created with information from (28).

From 2005 onwards, large-scale genome-wide association studies (GWASs) of obesity have aimed to uncover common SNPs, each with small additive or multiplicative effects, that together may explain a proportion of the heritability of the condition. From such analyses, over 300 SNPs associated with BMI, waist-to-hip ratio and other adiposity-related traits have been identified (35). Among the more robust and well-replicated findings have been the association between BMI and SNPs in the fat mass and obesity associated (*FTO*) gene. Several of these SNPs are located in the first intron of the *FTO* gene, and each additional minor (risk) allele is associated with a 0.39 kg/m2 higher BMI (36). Other obesity associated SNPs have more modest effect sizes of 0.06-0.33 kg/m^2 per BMI-increasing allele, and SNPs identified to date can only explain a small proportion of the variance in BMI. For example, a meta-analysis of BMI-associated SNPs identified through GWASs and Metabochips (custom genotyping arrays for genetic studies of metabolic, cardiovascular and

anthropometric traits) that included almost 340,000 individuals, identified approximately 100 loci, but these loci together accounted for only around 2.7% of variance in BMI (37). However, the variance explained by genetic known variants will likely increase; simulations based on whole-genome sequencing indicate that SNPs are able to explain approximately 27% of BMI variance (38).

The genetic background of T2DM is similar to that of obesity. Thus, whilst a small proportion of individuals develop the disease as a result of rare variants with high penetrance, the majority of cases occur as a result of several variants, each with small effect sizes (39).

One form of monogenic diabetes is maturity-onset diabetes of the young (MODY). This non-insulin dependent form of diabetes accounts for between ~ 1 and 4% of all cases of diabetes in those diagnosed under the age of 30 years, and is inherited in an autosomal dominant fashion (40). The disease can be caused by a mutation in one out of a number of different genes, several of which are involved in glucose and insulin signalling and function (41)(Figure 1-6). The majority of MODY cases arise as a result of a mutation in one of four genes: Glucokinase (*GCK*), Hepatocyte Nuclear Factor 1 Alpha (*HNF1*α), Hepatocyte Nuclear Factor 1 Beta (*HNF1*β) and Hepatocyte Nuclear Factor 4 Alpha (*HNF4*α) (42). Approximately one third of cases of MODY occur as a result of heterozygous mutations in *GCK*. Under normal circumstances, glucokinase acts as an intracellular glucose sensor such that as glucose enters pancreatic β cells, glucokinase catalyses the first step towards its conversion to ATP, which leads to downstream secretion of insulin (Figure 1-6). Mutations in *GCK* therefore reduce the potential of β cells to secrete insulin in response to glucose, resulting in fasting hyperglycaemia (41).

The hepatocyte nuclear factors are transcription factors that have complex and wide-encompassing roles in embryonic development and regulation of adult metabolism (43). They are expressed in a variety of adult human tissues, in particular the liver, kidney and pancreas (43). Among other cellular functions, *HNF1*α regulates transcription of insulin (*INS*) and glucose transporter 2 (*GLUT2*) a transmembrane

glucose carrier, in mature pancreatic β-cells (44). In line with this, *HNF1α* knock-out mice develop diabetes as a result of decreased insulin secretion (45). In humans, MODY caused by *HNF1α* mutations presents as fasting hyperglycaemia resulting from progressive β-cell dysfunction (43). *HNF4α* are thought to act in the same transcription pathway as HNF1α, and has a similar clinical phenotype as MODY caused by HNF1α mutations (43). MODY caused by *HNF1β* mutations, in contrast, is characterised both by β-cell dysfunction and insulin resistance, and commonly involve renal complications such as renal cysts (46). *HNF1β* is regulates transcription of *HNF1α* as well as *GLUT2*, exemplifying the interconnectedness of these key metabolic transcription factors (46).

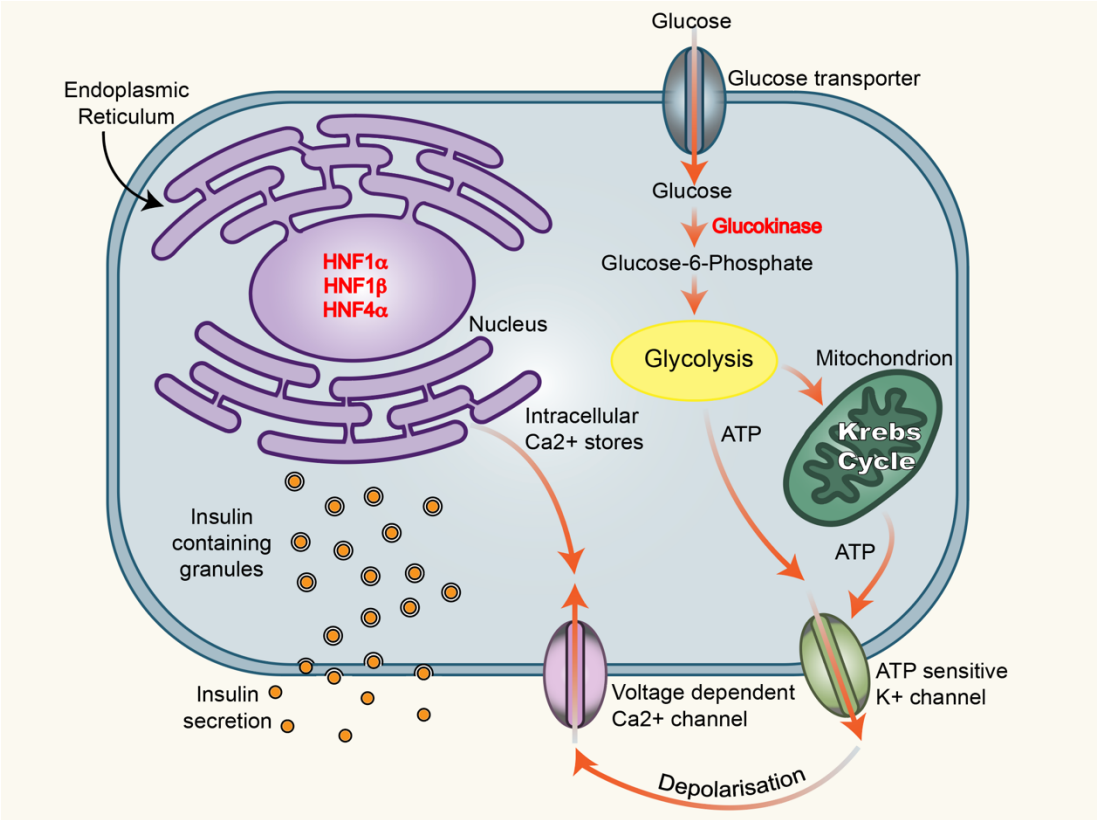


Figure 1-6. Schematic of a pancreatic β cell and genes implicated in Maturity Onset Diabetes of the Young (MODY).
Genes implicated in MODY are depicted in red.

The majority of cases of adult T2DM, however, have a polygenic basis, and numerous GWASs have attempted to uncover common genetic variants that through additive effects increase the risk of developing the condition (47-50). A recent study aggregated findings from 32 genome wide association studies of T2DM, allowing assessment of disease risk conferring genetic variants in almost 900,000 individuals, 8% of which had T2DM (51). The greatly expanded sample size compared to the individual GWASs, combined with high-density imputation led to the identification of 243 genome-wide significant loci, including rare (MAF<0.5%) but highly penetrant genetic variants with odds ratios as high as 8.05 ($p<5x10^{-8}$) (51). Despite the large sample size and sophisticated bioinformatic approaches for detecting causal genetic variants, however, the combined effect of the identified loci could only explain approximately 18% of T2DM risk (51).

In summary, the genetic contribution to obesity and T2DM is one in which a minority of cases result from rare genetic variants with high penetrance, whereas most cases

are due to several genetic variants, each with small effect sizes. However, there is at present still a considerable discrepancy between the relatively high heritability estimates for obesity and T2DM and the risk for these conditions attributable to known genetic variants (24). There are several potential underlying reasons for this "missing heritability". One is that we are at present unable to detect rare genetic variants, in particular those in regions of low linkage disequilibrium (23). Another is our yet limited ability to assess the influence of gene-environment interactions on phenotypic traits. As study sample sizes increase and bioinformatic approaches improve, the genetic contributions to obesity and T2DM will become increasingly better characterised. However, in parallel to deciphering the underlying genetic landscape of these conditions, there is also value in understanding the developmental, gene-environment and environmental factors that contribute towards the disease phenotypes.

1.3 Birth Weight and Adult Metabolic Health

An individual's risk of developing metabolic disease is influenced by their growth *in utero* (52, 53). It is well established that large for gestational age (LGA) infants, usually defined as a birth weight above the 90th centile for gestational age, have a higher risk of developing obesity and type 2 diabetes in later life (53). A meta-analysis of 20 studies investigating the association between birth weight and adult obesity risk showed that infants born LGA, here defined as a birth weight > 4000 grams, were at significantly higher risk of developing obesity in adolescence and adulthood when compared to infants with a birth weight < 4000 grams (OR 2.07; 95% CI 1.91-2.24) (53).

Maternal obesity and gestational diabetes mellitus (GDM) are major risk factors for giving birth to an LGA infant. Indeed, maternal obesity doubles the risk of LGA offspring (54). Gestational diabetes is loosely defined as any degree of glucose intolerance with its onset, or first diagnosis, during pregnancy (55). It usually resolves after delivery, but constitutes a risk factor for later development of T2DM in the mother (54). GDM affects approximately 5% of pregnancies worldwide, although this figure depends on maternal ethnicity, phenotype and the criteria for blood glucose

thresholds following a glucose tolerance test (GTT) (54). Unless adequately controlled, GDM is also significantly associated with increased offspring birth weight as well as an increased risk of diabetes in the next generation (56, 57). The mechanism behind this is thought that the increased glucose availability from the mother promotes fetal hypersecretion of insulin, which leads to fetal macrosomia (58, 59).

At the other end of the birth weight spectrum, small for gestational age (SGA) infants are also at risk of developing metabolic syndrome in adulthood. In the late 1980's, Barker and Hales noted a correlation between low birth weight and later development of lifestyle-related disorders, including impaired glucose tolerance, type 2 diabetes (T2DM), hypertension and cardiovascular disease (52). A key observational study was the Hertfordshire cohort study, which assessed birth weight records of 468 men born between 1920 and 1930 and found that the prevalence of impaired glucose tolerance, overt diabetes and hypertension at age 57-70 years fell progressively with higher birth weights such that the rate of T2DM was 40% if born at less than 2.5kg and less than 14% if born at more than 4.3kg (52). These and other findings led to the development of the 'thrifty phenotype' hypothesis (60). This hypothesis proposed that prenatal malnutrition leads to lifelong changes in insulin sensitivity, which together with lifestyle factors predisposes an individual to components of the metabolic syndrome in later life (60). The concept has been expanded to the "developmental origins of disease hypothesis" which broadly proposes that adult disease risk can be programmed by the perinatal environment (61).

Further evidence that low birth weight is associated with an increased risk of adult metabolic disorders was provided by the Dutch Hunger Winter, a historical disaster that nevertheless has improved our understanding of the link between fetal growth and adult health. The Dutch Hunger Winter describes a period of extreme and acute food shortage in the West Netherlands between December 1944 and May 1945. This event occurred due to a German blockade against food shipments, resulting in daily food rations of only 400-800 kcal per day (62). This provided a key setting to study

the association between prenatal undernutrition and adult health because of several unique characteristics; 1) the famine was imposed on a previously well-nourished population, 2) detailed prenatal and birth records were kept throughout the period and 3) when the famine abruptly ended the population went back to normal nutritional intake. Birth weight was affected by the stage of gestation at which fetuses were exposed to maternal famine; if affected in late pregnancy they were lighter than those affected in early gestation (where the mother went back to normal nutritional intake towards late pregnancy). Even so, early nutritional deprivation had lasting effects on adult health. At age 50 and 58 years, individuals who had been exposed to famine early in pregnancy had a higher average BMI, a more atherogenic lipid profile and a higher prevalence of cardiovascular disease than unexposed individuals. Exposure to famine at any stage of gestation was associated with impaired glucose tolerance. Low birth weight was specifically associated with hypertension (62).

More recently, large-scale meta-analyses have supported the association between low weight at birth and the risk of adult metabolic syndrome (63). A recent meta-analysis that included a total of 7,646,267 participants showed that for each kilogram increment in birth weight, there was a 22% reduction in risk of later developing T2DM (OR ratio: 0.78, 95% CI: 0.70–0.87) (63) (Figure 1-7). The same study showed in a binary analysis that participants with a birth weight <2.5 kg experienced a 45% (OR: 1.45, 95% CI: 1.33–1.59) higher risk of T2DM than those with a birth weight ≥2.5 kg (63). In addition, each kilogram increment in birth weight was associated with a 16.5% reduction in risk of developing CVD (OR: 0.84, 95% CI: 0.81–0.86) (63).

Studies of the association between low birth weight and adult risk of obesity has yielded more mixed results. Some observational studies suggest that low weight at birth is associated with an increased prevalence of adult obesity. Thus, in a cross-sectional study of approximately 2,500 adults in the Swiss CoLaus cohort, a birth weight of ≤2.5 kg in women was associated with an increased risk of obesity in adulthood when compared to the reference birth weight category of 2.5-3.5 kg ($p < 0.001$). However, this study failed to correct for gestational age at birth (64). Large-

scale meta-analyses have contradicted such findings. For example, a meta-analysis by Zhao *et al.* that examined data from fifteen studies of low birth weight and adult obesity found no evidence for an association between a birth weight of <2.5 kg and adult obesity when compared to birth weights between 2.5-4 kg (OR = 1.17, 95% CI 0.94, 1.46) (65).

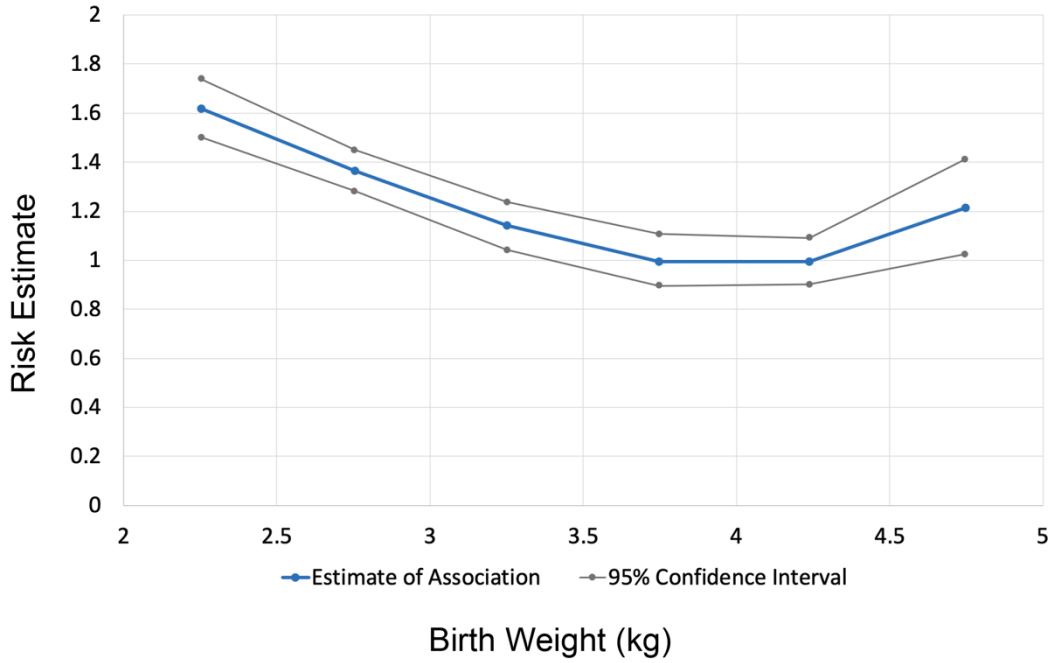


Figure 1-7. Dose-response relationship between birth weight and adult risk of T2DM from a meta-analysis including a total of over 4,000,000 participants.
The blue line represents point estimates of association, and the grey lines are the corresponding 95% CI. Data derived from (65).

Part of the explanation for the mixed results may reside in the impact of accelerated postnatal 'catch-up' growth, typically described as low birth weight infants gaining weight within their first two years of life such that their age-adjusted centile for weight is significantly higher at age 1 or 2 years compared to that at birth. Catch-up growth is now considered to be particularly detrimental for adult risk of metabolic disorders, pointing towards an increased 'metabolic vulnerability' of low birth weight infants towards lifestyle-related disease (66).

In summary, however, the consensus is that there is a U-shaped correlation between weight at birth and the risk of developing metabolic syndrome in adulthood, such that birth weights at both ends of the spectrum predispose to components of the metabolic syndrome.

1.4 The Paternal Influence on Offspring Birth Weight

Fetal growth *in utero* is determined by an interaction between the fetal genotype, where 50% of genes come from each parent, and the in-utero environment, which is wholly provided by the mother. As previously discussed, the intrauterine environment, for example the availability of glucose, is a crucial determinant of fetal growth. However, it has become increasingly evident that paternal factors also have an important influence on his baby's birth weight.

Magnus *et al.* examined the complete birth cohort of Norway 1967-98, which included nearly 70,000 mother-father-firstborn child trios, and demonstrated that paternal weight at birth was a significant determinant of offspring birth weight (67). There was an almost linear correlation between paternal birth weight and offspring birth weight within groups of maternal birth weight (

Figure 1-8). For example, if the mother had a normal birth weight of 3500-3999g but the father had a low birth weight of <2500g, the relative risk of their offspring being born with a low birth weight of <2500g was 2.0. If the father and mother had a normal birth weight of 3500-3999g, this risk drops to 1.1 (67). The authors conclude that there is a significant paternal genetic contribution towards offspring weight at birth, and suggest a heritability estimate for birth weight of approximately 0.25 (67).

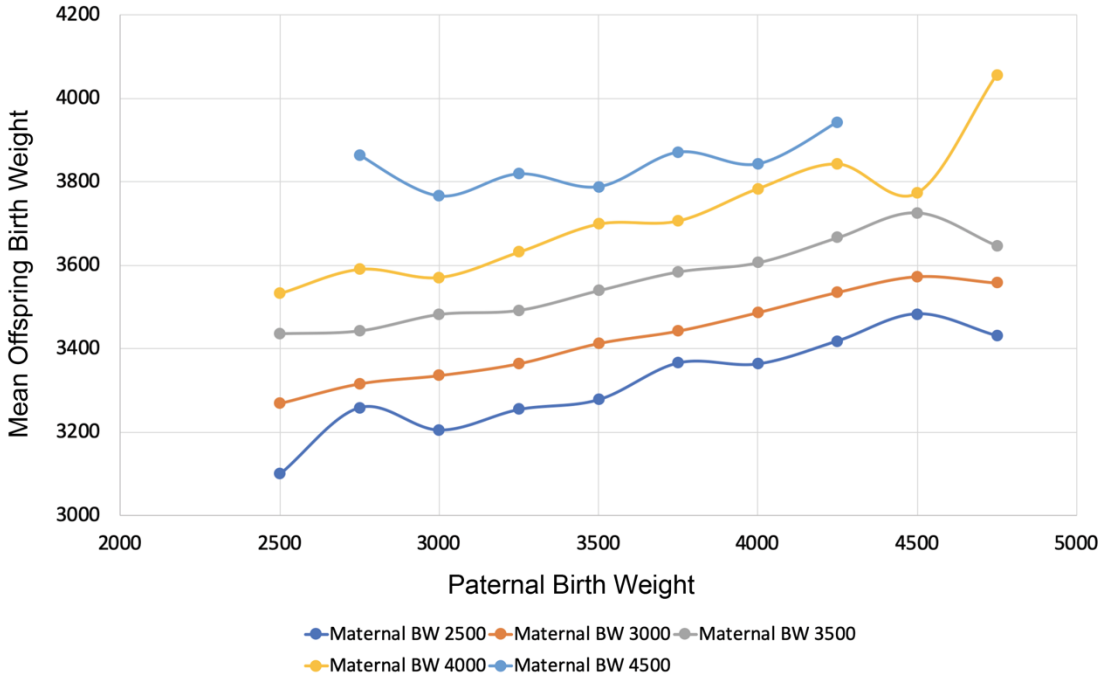


Figure 1-8. In a study of almost 70,000 mother-father-firstborn child trios, Per Magnus *et al.* found an almost linear increase in mean offspring birth weight according to paternal birth weight group. The differently coloured lines indicate groups of maternal weight at birth as detailed in the legend. Maternal and paternal birth weights are placed into 250g groups such that birth weight group 2500 = 2500-2749, 2750 = 2750-2999, 3000 = 3000-3249 etc. Data derived from (67).

More recent research has suggested that not only genetic, but also acquired paternal traits, including obesity and insulin resistance, have the potential to influence fetal growth and weight at birth. A nested cohort study by McCowan *et al.* involving 2002 couples in the prospective Screening for Pregnancy Endpoints (SCOPE) cohort, found that men who fathered SGA infants, defined as a birth weight <10th customised birth weight centile, were more likely to be obese compared with men who fathered normally grown infants (adjusted OR 1.50, 95% CI 1.05-2.16) (68). The study adjusted for maternal factors such as age, ethnicity, BMI, smoking status and maternal birth weight. Fathers of SGA offspring were also more likely to have central adiposity, defined as a waist circumference >102 cm (68). In line with the study by Magnus *et al.* discussed above, men who fathered SGA infants were found to have had a lower birth weight themselves (mean 180g lighter, $p < 0.0001$), supporting a self-perpetuating cycle of low birth weight, adult obesity and an increased risk of fathering SGA offspring.

A study by Hillman *et al.* provided further support for the association between paternal metabolic syndrome and low birth weight offspring. This case-control study compared metabolic parameters of men who had recently fathered pregnancies affected by SGA (n = 42, mainly FGR) with men who fathered appropriately grown neonates (n = 77). Insulin resistance was determined by comparing homeostatic model assessment of insulin resistance (HOMA-IR) values, which evaluates insulin resistance based on fasting blood glucose and insulin levels, between the two groups of fathers. Fathers of SGA offspring were more likely to be sub-clinically insulin resistant (OR 7.68 of having a 1 unit higher log HOMA-IR value; 95% CI 2.63–22.40; p = <0.001), and were also more likely to smoke (OR 3.39; 95% CI 1.26–9.16; p = 0.016), when controlling for factors such as maternal disease, age, BMI, ethnicity, and parity (69). In line with previous research discussed above, this study also found that fathers of SGA offspring were more likely to themselves have been light at birth (birth weight 3127±597g for fathers of SGA offspring, versus 3506±380g for fathers of AGA offspring, p = 0.0045).

Observational studies therefore suggest that whereas maternal obesity and insulin resistance is associated with LGA offspring, paternal metabolic syndrome may be associated with an increased risk of fathering SGA offspring (65, 68, 69).

1.5 The Genetic Association between Paternal Insulin Resistance and Offspring Birth Weight

Clues towards the mechanism behind the epidemiological association between paternal insulin resistance and low offspring birth weight may be provided by the influence of forms of monogenic diabetes on the birth weight of an infant whose father has the condition. Indeed, *GCK* mutations in MODY provide an example of a genetic variant that confers insulin resistance in a father as well as predisposes his offspring both to low weight at birth and insulin resistance in adulthood (70).

As discussed in section 1.2, glucokinase (*GCK*) mutations lead to defective glucose sensing and insulin secretion. Women heterozygous for a *GCK* mutation are typically asymptomatic until screened for gestational diabetes, when they are found to have fasting hyperglycaemia. This glucose overload stimulates fetal insulin secretion and women with a *GCK* mutation, therefore, tend to give birth to LGA infants, provided that the fetal genotype is normal. If there is ultrasound evidence of macrosomia in pregnancies of women with *GCK* mutations, therefore, they should be offered insulin treatment during pregnancy (71). The effect is large; infants born to women with *GCK* heterozygosity are on average 601g heavier at birth than AGA infants ($p = 0.001$) (72). In cases where a fetus is heterozygous for a *GCK* mutation and the mother has a normal genotype, i.e. where the fetus has inherited the condition from the father or developed it *de novo*, the fetus will fail to secrete enough insulin in response to glucose, and will be SGA. Indeed, the average birth weight of these infants is decreased by as much as 533g ($p = 0.002$) (72). If both the mother and the fetus are heterozygous for a *GCK* mutation, the effects of the hyperglycaemic environment and the low insulin secretion in the fetus effectively cancel each other out, resulting in normal birth weight infants.

In summary, there are genetic variants that both predispose a fetus to poor intrauterine growth and to an increased risk of adult diabetes. This concept is known as the fetal insulin hypothesis (70).

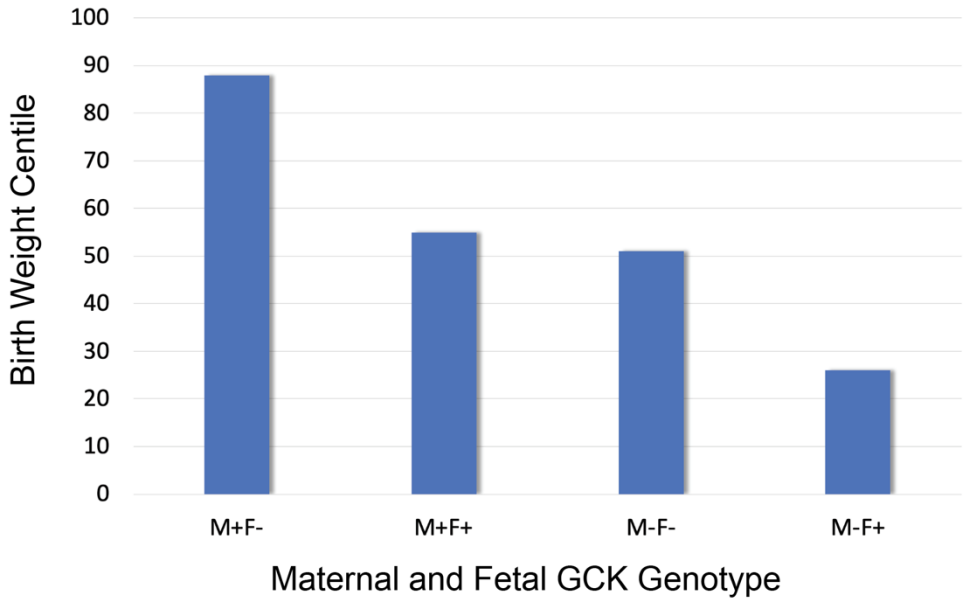


Figure 1-9. Birth weight centile distribution in 58 offspring in families with *GCK* mutations according to maternal and fetal *GCK* genotype.
M+F- denotes a maternal *GCK* mutation and a normal fetal genotype, M+F+ denotes a maternal and a fetal *GCK* mutation, M-F- denotes a normal maternal and fetal genotype and M-F+ denotes a normal maternal genotype and a fetal *GCK* mutation. Figure created using data from (70).

Single gene mutations like MODY provide an example of the link between parental insulin resistance and offspring growth. However, most cases of diabetes are a result of additive effects from a large number of genetic variants, each with small effect sizes (section 1.2). From this follows that there would be an overlap between genetic variants identified in large cohorts of individuals with T2DM and genetic variants identified in large cohorts of low birth weight infants. A GWAS meta-analysis of birth weight in over 150,000 individuals identified 60 SNPs that were significantly associated with birth weight at a genome-wide level ($p < 5 \times 10^{-8}$) (73). Nine of these had previously been associated with adult T2DM (73). However, taken together, the 60 birth weight loci could only explain approximately 2% of variance of birth weight. This indicates either that increasingly large sample cohorts are required to detect underlying genetic effects, or that other factors, such as gene-environment interactions, underlie part of the epidemiological association between low weight at birth and an increased risk of metabolic syndrome in adulthood (73, 74).

1.6 Bridging the Gap: Epigenetic Inheritance and Fetal Growth

At present, only a small proportion of the epidemiological association between paternal metabolic disorders, e.g. obesity and T2DM, and fetal growth can be explained by genetic associations. Although improvements in methodology and increased sample sizes are likely to identify a larger number of genetic variants predisposing to both dysregulated fetal growth and adult risk of diabetes in the years ahead, the strong impact of environmental and lifestyle-related factors on the pathogenesis of T2DM makes a plausible case for gene-environment, or epigenetic, changes, to explain part of the impact of paternal health on fetal growth.

Epigenetics is generally defined as 'the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence' (75, 76). In other words, epigenetics refers to regulation of gene expression that do not involve changes in the underlying genetic code. As such, while the genetic code is preserved in all human cell types, epigenetic alterations allow gene expression to be regulated in a tissue- cell- and developmental stage specific manner (77). Epigenetic changes are dynamic, allowing organisms to alter their gene expression in response to factors such as environmental conditions, tissue specific requirements and nutrient availability (78, 79). In broad terms, epigenetic mechanisms act by altering the accessibility of chromatin towards the transcriptional machinery, and by regulating gene expression at a post-transcriptional level (80, 81). There are three distinct but interrelated epigenetic mechanisms (Figure 1-10):

(1) DNA methylation: the addition of a methyl group to the 5-carbon of cytosine, forming 5mC,
(2) posttranslational modifications of histones, including methylation, acetylation, phosphorylation and sumoylation, and
(3) noncoding RNAs (ncRNAs)

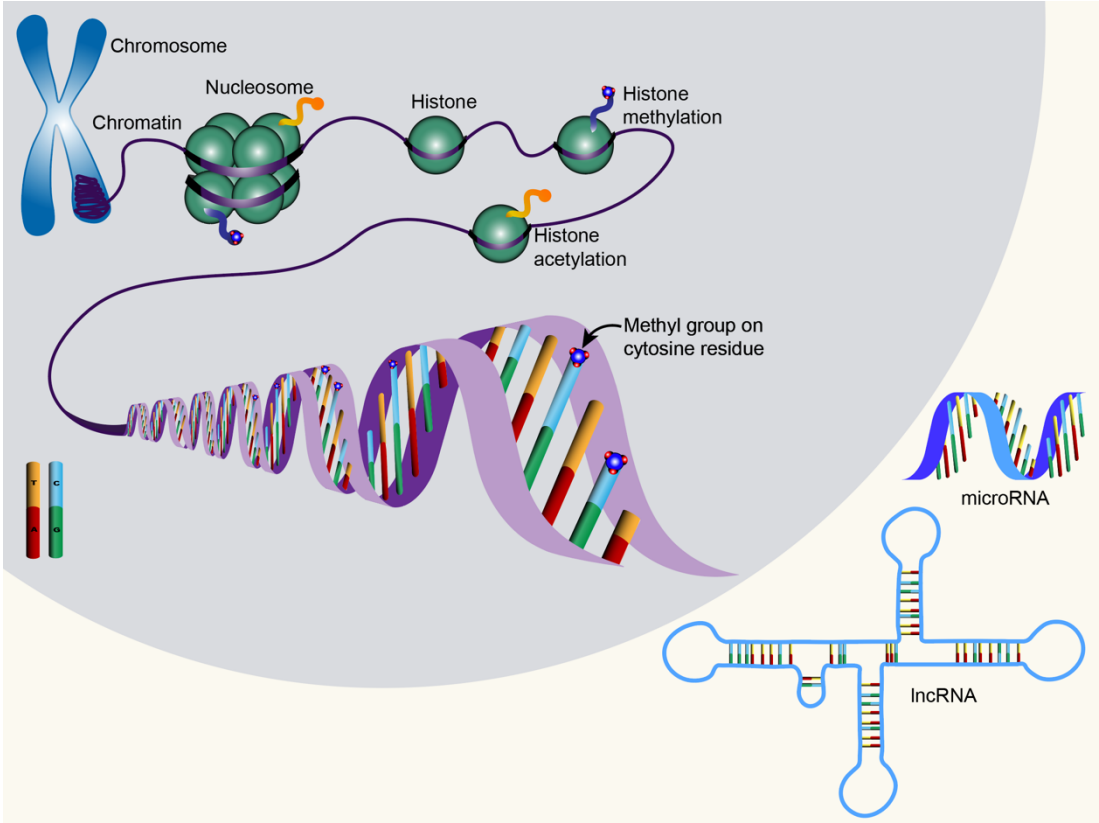


Figure 1-10. Overview of epigenetic mechanisms.
Epigenetics is generally taken to include three distinct but interrelated mechanisms; DNA methylation, posttranslational modifications of histones, and non-coding RNA.
lncRNA = long non-coding RNA.

(1) Methylation of mammalian genomes occurs predominantly at cytosines next to guanines ('CpG sites'). DNA methylation at gene promoters and enhancers is generally associated with transcriptional silencing, whereas DNA methylation in the gene body tends to be associated with active gene expression, although there are exceptions (79). Removal of DNA methylation can occur both by passive and active mechanisms. Passive demethylation of DNA occurs in the absence of maintenance methylation of newly synthesised DNA strands during replication (82). The active process involves enzymes such as ten eleven translocation (TET) enzymes. TET enzymes can oxidise 5mC to yield 5-hydroxymethylcytosine (5hmC) (82). Further oxidation, again by TET enzymes, generates 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (82). 5fC and 5caC can be cleaved by thymine-DNA glycosylase (TDG) and then recognised by the base excision repair system in which they are replaced with an unmethylated cytosine (82).

(2) Histone methylation can either repress or activate transcription, depending on which lysine is methylated. For example, trimethylation of histone H3 at lysine 4 (H3K4me3) is associated with active gene transcription, whereas dimethylation of histone H3 at lysine 9 (H3K9me2) is associated with transcriptional silencing (83). Unlike methylation, acetylation of histones potentiates and deacetylation suppresses gene expression (83)

(3) Non-coding RNAs, including microRNAs (miRNAs), PIWI-interacting RNAs (piRNAs), and long non-coding RNAs (lncRNAs), can affect RNA silencing and post-transcriptional regulation of gene expression and be transmitted independently of the genetic sequence and are thus also considered part of the epigenetic machinery (84).

1.7 Overview of Epigenetic Profiling

Regulation of epigenetic processes is fundamental to normal mammalian development and its dysregulation has been linked to a wide range of disease processes (79). Thus, the last couple of decades have seen increasingly sophisticated methods of characterising epigenetic signatures associated with a range of human traits and diseases. Characterisation of epigenetic signatures of disease can aid the understanding of disease pathogenesis, develop biomarkers for disease progression and identify therapeutic targets (85). The majority of studies thus far have focussed on DNA methylation, which will also form the focus of the methods of epigenetic profiling discussed below.

1.7.1 Overview of DNA Methylation Profiling

DNA methylation is fundamental in biological processes such as X chromosome inactivation, silencing of retroviral transposable elements and long-term regulation of gene expression (86). DNA methylation is the most studied epigenetic mark in human health and disease to date (79). This is partly owing to it being more stable than non-coding RNA and histone modifications, and to the availability of commercial, cost-effective methylation analysis platforms that allow single base resolution profiling of targeted CpG sites using probes on a microarray.

The current gold standard technique for analysing DNA methylation is whole-genome bisulphite sequencing (WGBS). This technique commences with the treatment of genomic DNA with sodium bisulfite, which converts unmethylated cytosines (C) to uracil (U), while leaving methylated cytosines unchanged. During subsequent whole genome amplification, uracil is amplified to its complementary base thymine (T). These steps effectively transform the DNA methylation analysis into a genotyping assay, as the initially methylated CpG sites can be identified through determining the presence of C versus T. Bisulfite treatment and whole genome amplification is followed by whole-genome highly parallel sequencing.

WGBS has been successfully applied to create a complete map of the ~ 28 million CpG sites in the human genome using a variety of tissues and cell types (87). However, due to the high cost per sample and expertise required to analyse WGBS data, this technique is not always the most practical. Thus, microarray-based technologies that enable the interrogation of a large number of DNA fragments in a highly parallel fashion have become widely used and have allowed the completion of large-scale epigenome-wide association studies into multiple complex human conditions. Among microarray-based technologies, the lllumina BeadChips are among the most extensively employed, and consist of three-micron silica beads on a solid surface, each bead covered with hundreds of thousands of copies of a specific oligonucleotide, which act as the capture sequences in a given Illumina assay.

The first Illumina Methylation BeadChip, the HumanMethylation27K (HM27), was introduced in 2008. The array interrogated methylation across >25,000 CpG sites, mainly within promoter regions of well-described genes. The HM27 allowed the first epigenome-wide association studies (EWASs) to be performed, for example leading to an improved understanding of the role of aberrant DNA methylation in carcinogenesis and its association to ageing and smoking (88, 89).

The HM27 was superseded by the Illumina HumanMethylation450 (HM450) in 2011. The HM450 allowed investigation of >485,000 CpG methylation sites across the genome, including CpG islands, shores and shelves, 5’ UTR, 3’ UTR and bodies of

RefSeq genes, the MHC region and some enhancer regions (87). The HM450 represented a significant improvement compared to its predecessor, and rapidly became the most widely used platform to study DNA methylation and its association to a broad range of conditions, including several forms of cancer, immune related disorders such as rheumatoid arthritis and the effect of maternal smoking in pregnancy on the neonate (90-93).

However, the HM450 was not without issues. For example, in the years since the HM450 was introduced, the importance of DNA methylation in the regulation of transcriptional activators binding to genomic enhancer regions was increasingly recognised (87). It was therefore problematic that CpGs from only a small proportion of enhancer regions were represented on the array (86). Enhancer regions, which may be located hundreds of thousands base pairs away from the transcription start site, allow transcriptional regulation by binding to transcriptional activators, which enhance the activity of the promoter e.g. by stabilising interactions between transcription factors and the transcription machinery (Figure 1-11)(86).

The most recent of the Illumina DNA methylation arrays is the Illumina MethylationEPIC Array (EPIC), which was introduced in 2015. The EPIC array was specifically designed to interrogate potential enhancer regions as well as 90% of the CpG sites assayed by the HM450 (79). The Illumina MethylationEPIC Array is described in detail in Chapter 2 section 2.12.

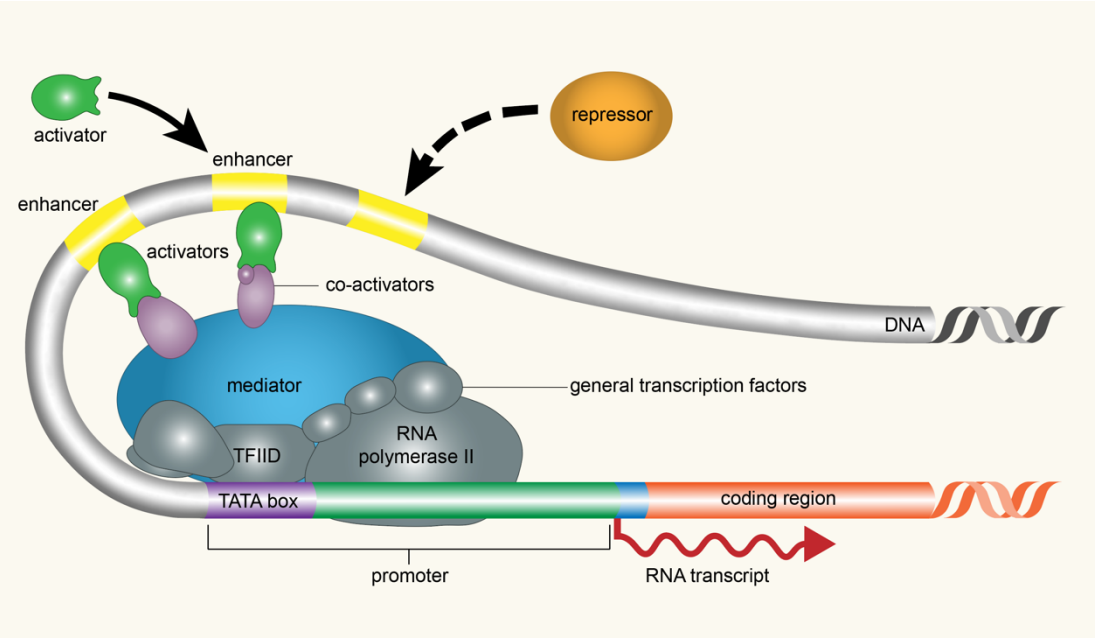


Figure 1-11. Transcriptional regulation via enhancer regions.
TATA box = genomic region containing a high density of T and A and which forms part of the promoter region. TFIID = Transcription Factor II D.

1.7.2 Overview of Histone Modification Profiling

There is a range of methods for detecting targeted, global and genome-wide post-translational modifications (PTMs) of histones. Among the most widely used methods of histone profiling is chromatin immunoprecipitation (ChIP). Briefly, this involves cross-linking DNA-histone interactions with formaldehyde followed by fragmentation of the DNA (bound to histones). Antibodies are then targeted towards the specific histone modification of interest and the DNA bound to the isolated histone can be analysed e.g. by quantitative PCR (ChIP-PCR), microarray-hybridisation (ChIP-chip) or next generation sequencing (ChIP-seq) (94).

More recently, issues such as antibodies cross-reacting with similar modifications on the same histone protein, or on a different histone protein, as well as the requirement for *a prioi* knowledge about the type of modification of interest, have made approaches such as mass spectrometry more widely used for studying histone modifications. Mass spectrometry, in contrast to antibody-based methods, allows unbiased profiling of several PTMs simultaneously (95).

1.7.3 Overview of Profiling non-coding RNA

Non-coding RNA species are RNA species that do not encode for proteins. They constitute a diverse set of RNA molecules which includes long non-coding RNAs (lncRNAs), transfer RNAs (tRNAs), ribosomal RNAs (rRNAs) as well as several types of small RNA species such as microRNAs, small interfering RNAs (siRNAs) and piwi-interacting RNA (piRNAs).

The gold standard for RNA profiling is next generation RNA sequencing. This method is replacing previously popular hybridization-based microarray methods, and allows an unbiased assessment of the presence and quantity of RNA molecules in biological samples (96). In brief, the method involves extracting messenger RNA (mRNA) from the organism, tissue or cell of interest, fragmenting the mRNA, and generating double stranded complementary DNA (cDNA) using reverse transcriptase. The cDNA molecules are then amplified, fragmented and sequenced based on a reference genome. Aligning the RNA-seq reads against a reference genome also allows mapping them into genomic positions. For sequencing of small RNA species, the cellular RNA is size selected prior to generation of cDNA, e.g. with a size exclusion gel of size selection magnetic beads.

1.8 Epigenetics and Metabolic Syndrome

The rapid rise in the prevalence of obesity cannot be explained by genetics alone, but more adequately by a combination of genetic susceptibility, environmental factors and gene-environment interactions. This observation has spurred intensive efforts into uncovering which, if any epigenetic variants are associated with the disease phenotypes (24).

In non-human mammals, there are convincing examples of how specific epigenetic perturbations can not only be associated with metabolic phenotypes, but also be causal (97). For example, knockout of the H3K9-specific demethylase Jhdm2a leads to obesity and hyperlipidaemia in mice (98). Mechanistically, it was shown that that this single epigenetic factor deficiency decreased the physiological ability to respond

to β-adrenergic stimulation via two key regulators of metabolism; uncoupling protein 1 (UCP1) and peroxisome proliferator-activated receptor alpha (PPARα) (98).

In humans, however, intervention studies are less feasible from an ethical and methodological perspective. Instead, most studies have been increasingly large-scale epigenome-wide association studies (EWASs) of DNA methylation in obesity and T2DM. The underlying hypothesis has been that the epigenetic architecture of these conditions is equivalent to their genetic counterpart, i.e. that in the majority of cases, several epigenetic variants, each with small effect sizes, contribute towards the phenotype. In the discussion below, I have specifically focused on DNA methylation signatures of obesity as opposed to those of T2DM. This was partly to limit the scope of the discussion, but also because the review of studies investigating DNA methylation in obesity formed the basis of a validation experiment aimed to generate a robust and reproducible DNA methylation profile described in Chapter 6.

1.8.1 DNA Methylation in Human Obesity

Early studies of DNA methylation in human obesity frequently focussed on candidate genes. These included imprinted genes with known involvement in growth regulation, e.g. *IGF2/H19*, and genes with known functions in appetite regulation, e.g. pro-opiomelanocortin (*POMC*) (99). However, such hypothesis-driven approaches yielded inconsistent results. Since the introduction of array-based DNA methylation analyses tools, studies have progressed to a less biased interrogation of the methylome as well as to include increasingly large sample cohorts. In order to summarise the current understanding of the association between DNA methylation and human obesity, I performed a review of studies using the PubMed database in June 2019. The following selection criteria were employed:

- Analysis of DNA methylation in association to human obesity or BMI
- Minimum sample size of 1000 participants (discovery and replication cohorts combined)
- Performed using the Illumina HM450 array (to make results comparable across studies)

Nine studies were identified using the above criteria and are summarised in Table 1-1. All studies analysed blood or blood components. The minimum cut-off of 1000 participants was chosen to limit the scope of the review, and because the robustness of findings from EWASs is closely related to its sample size. Indeed, among the studies included in this review, there was a clear correlation between the size of the discovery cohort and the number of genome-wide significant CpG sites identified in the EWAS (R^2 = 0.765, p = 0.001) (Figure 1-12).

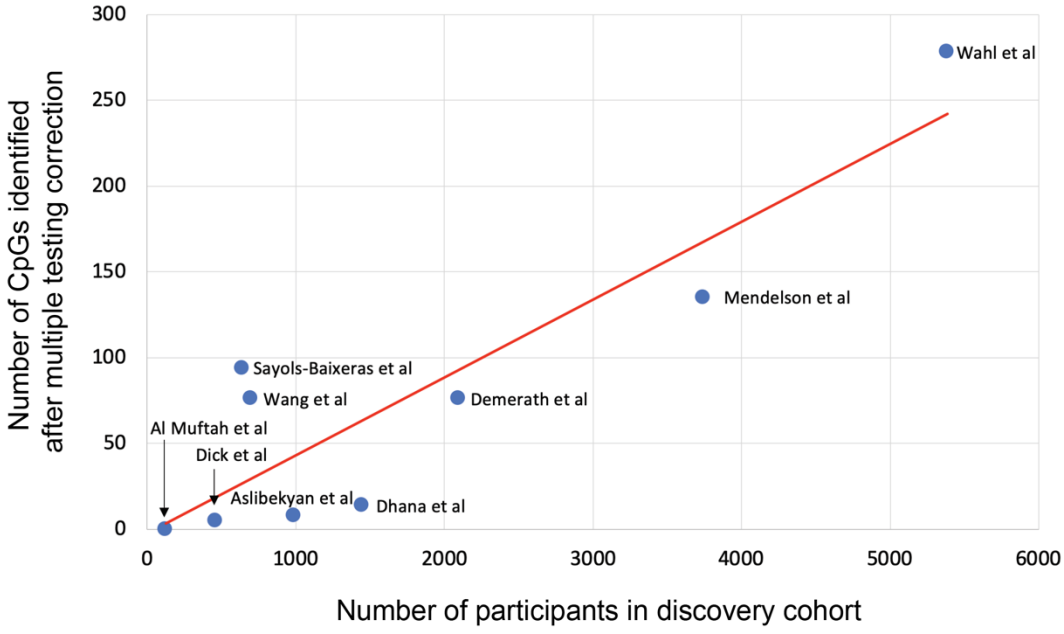


Figure 1-12. Correlation between number of participants in obesity EWAS discovery cohorts and number of CpG sites identified after correcting for multiple testing.
R^2= 0.765, p = 0.001. For references to the original publications see Table 1-1.

Author, year, reference		*n*	*Phenotype*	*Tissue*	*Number of CpG Sites identified (multiple testing correction)*	*Ethnicity*	*Cohort*
Dick et al., 2014 *(100)*	Discovery cohort	459	BMI	Whole blood	5 (FDR)	Caucasian	Cardiogenics Consortium
	Replication cohort	339	BMI	Whole blood	3 (Bonf.)	Caucasian	MARTHA
	Replication cohort 2	1789	BMI	Whole blood	3 (Bonf.)	Caucasian	KORA
Aslibekyan et al., 2015 *(101)*	Discovery cohort	991	BMI	CD4+ T cells	8 (Bonf.)	European Americans	GOLDN
			WC		5 (Bonf.)		
	Replication cohort 1	2105	BMI	Whole blood	2 (Bonf.)	European Americans	FHS
			WC		3 (Bonf.)		
	Replication cohort 2	1935	BMI	Whole blood	4 (Bonf.)	African Americans	ARIC
	Meta-analysis	5031	BMI	Whole blood	8		
		3096	WC		5		
Demerath et al., 2015 *(102)*	Discovery cohort	2097	BMI	PBLs	76 (Bonf.)	African Americans	ARIC
			WC		164 (Bonf.)		
	Replication cohort	3368	BMI	Whole blood/ CD4+ T cells	37 (Bonf.)	European Americans	FHS + GOLDN
		991	WC	CD4+ T cells	8 (Bonf.)	European Americans	GOLDN
Al Muftah et al., 2016 *(103)*	Discovery cohort	123	BMI	Whole blood	None	Arab	Quatari family study
			T2DM		1 (Bonf.)		
	Replication cohort	810	BMI	Whole blood	None	Caucasian	TwinsUK
			T2DM		None		
	Meta-analysis*	123 + 810	BMI	Whole blood	None		
			T2DM		None		

Mendelson et al., 2017 *(104)*	Discovery cohort	3743	BMI	Whole blood	135 (Bonf.)	European Americans + Caucasian	FHS + LBCs
	Replication cohort	4055	BMI	Whole blood/CD4+ T cells	83 (Bonf.)	African Americans + European Americans + Caucasian	ARIC + GOLDN + PIVUS
Sayols-Baixeras et al., 2017 *(105)*	Discovery cohort	641	BMI	Whole blood	94** (Bonf.)	European	REGICOR
	Replication cohort	2515	WC	Whole blood	49** (Bonf.)	European Americans	FOS
Wahl et al., 2017 *(106)*	Discovery cohort	5387	BMI	Whole blood	278 (Bonf.)	European + Indian-Asian	EPICOR + KORA + LOLIPOP
	Replication cohort	4874	BMI	Whole blood	187 (Bonf.)	European + Indian-Asian	ALSPAC + EGCUT + Leiden Longevity + LifeLines Deep + LOLIPOP + RS-BIOS + RS-III + TwinsUK
Dhana et al., 2018 *(107)*	Discovery cohort	1450	BMI	Whole blood	14 (Bonf.)	European	Rotterdam Study
			WC	Whole blood	26 (Bonf.)		
	Replication cohort	2097	BMI	Whole blood	12 (Bonf.)	African Americans	ARIC
			WC	Whole blood	13 (Bonf.)		
Wang et al., 2018 *(108)*	Discovery cohort	700	BMI (obese vs lean)	PBLs	76 (Bonf.)	African Americans (14-36 years)	EpiGO, LACHY, BP Stress Study
	Replication cohort	2097	BMI	PBLs	54 (FDR)	African Americans	ARIC
	Replication cohort	188	Obese vs lean	Neutrophils	37 (FDR)	African Americans	Subgroup of EpiGO

Table 1-1. EWASs of obesity associated genome-wide CpG methylation.
Legend continued on the following page.

Table 1.1. Analyses of DNA methylation associated with obesity

* Study replicated 1 out of 8 T2DM and 7 out of 39 BMI findings from previous EWASs in the Quatari discovery cohort. The meta-analysis was performed on these 8 replicated loci

** Results are from a meta-analysis of discovery and replication cohorts.

WC = Waist Circumference, PBL = Peripheral Blood Leukocyte, FDR = False Discovery Rate, Bonf = Bonferroni, ARIC = Atherosclerosis Risk in Communities, MARTHA = MARseille THrombosis Association Study, KORA = Cooperative Health Research in the Augsburg Region, GOLDN = Genetics of Lipid Lowering Drugs and Diet Network, FHS = Framingham Heart Study, PIVUS = Prospective Investigation of the Vasculature in Uppsala Seniors, REGICOR = Girona Heart Registry, EpiGO = Epigenetic Basis of Obesity-Induced Cardiovascular Disease and Type 2 Diabetes, FOS = Framingham Offspring Study, LOLIPOP = London Life Sciences Population Study, ALSPAC = Avon Longitudinal Study of Parents and Children, EGCUT = Estonian Genome Center of the University of Tartu, RS-BIOS = Rotterdam Study Bios Cohort, RS-III = Rotterdam Study-III, LACHY = Lifestyle, Adiposity, and Cardiovascular Health in Youth.

A general observation is that there is considerable overlap in the study cohorts that have been analysed in these EWASs. For example, the Atherosclerosis Risk in Communities (ARIC) cohort has been included as a discovery cohort in one study, but as a replication cohort in a further four studies. The Framingham Heart Study (FHS) cohort has been studied as a discovery cohort in one study and a replication cohort in a further two studies. It is therefore possible that findings that appear to be robust and reproducible partly arise from studies that sample the same cohorts as previous researchers have done.

The first large scale obesity EWASs was published in 2014 (100). The study identified 3 CpG sites that were associated with BMI on a genome-wide level in both the discovery and replication cohorts (100). Interestingly, all three of these probes mapped to intron 1 of Hypoxia Inducible Factor 3 Subunit Alpha (*HIF3A*), a transcription factor that is involved in regulating cellular responses to hypoxia. The same study found that one of the significant probes, cg22891070, was also associated with BMI in adipose tissue, but not in skin ($p = 1.72 \times 10^{-5}$ and $p = 0.882$ respectively). Demerath *et al.* similarly identified an association between methylation of *HIF3A* and BMI and waist circumference in a cohort of over 2000 adults (102). Notably, these studies included cohorts with different ethnic backgrounds as discovery cohorts, suggesting that the association may be relatively independent of underlying genetic architecture. Subsequent research has supported the involvement of *HIF3A* methylation in the development of obesity. Thus, Pfeiffer at al showed that *HIF3A* methylation at cg22891070 was significantly higher in visceral adipose tissue compared to subcutaneous adipose tissue, and that its methylation level correlated with BMI (109).

There is, however, an overall lack of concordance of findings. The two largest studies to date found 187 obesity-related CpGs (Wahl et al) and 83 CpGs (Mendelson et al), but only 37 CpGs overlapped between the two studies (104, 106). Discrepancies could in part be explained by studying populations of different ethnicities or ages. However, in the context of deciphering disease pathophysiology and identifying

biomarkers for disease progression, the current lack of a robust, reproducible epigenetic signature of obesity presents a hindrance.

Despite the paucity of concordance between some study results, there are individual findings that have been convincingly replicated. These include CpG sites annotated to carnitine palmitoyltransferase 1A (*CPT1A*). In a study of 991 participants in the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) cohort, Aslibekyan *et al.* found that methylation of intron 1 of *CPT1A* was significantly associated both with BMI and waist circumference (101). This finding was replicated in both a Quatari and a UK cohort (103). Further, Demerath *et al.* identified *CPT1A* methylation as associated with BMI and waist circumference, and showed that that *CPT1A* methylation is associated with BMI in adipose tissue (102). Similarly, in the largest obesity EWAS performed to date, Wahl *et al.* found that *CPT1A* methylation is significantly associated with BMI (106). With the caveat that some of these studies had partially overlapping study cohorts, it thus seems that *CPT1A* methylation is robustly associated with BMI and related traits. The gene product of *CPT1A*, carnitine palmitoyltransferase 1A, is involved in the transport of fatty acids across the mitochondrial inner membrane (102). Indeed, deficiency of carnitine palmitoyltransferase results in a decreased rate of fatty acid beta-oxidation (110). Therefore, epigenetic dysregulation of *CPT1A* appears to be involved in the pathogenesis of obesity (103).

Another significant and replicated finding from obesity EWASs is differential methylation of probes annotated to Sterol Regulatory Element Binding Transcription Factor 1 (*SREBF1*). This transcription factor, sometimes called the 'master regulator of lipid homeostasis', regulates expression of the low density lipoprotein receptor and other genes involved in cholesterol synthesis (102, 111). Multiple studies have shown that individuals with obesity have differential DNA methylation at CpG sites associated with *SREBF1* (102-107). Dysregulated expression of the *SREBF1* gene product has consistently been linked to conditions such as non-alcoholic fatty liver disease (NAFLD) and insulin resistance in humans and mice (112, 113). In a large-scale meta-analysis of GWASs investigating T2DM, which included over 70,000 individuals,

independent SNPs at *SREBF1* reached study-wide significance ($p<2.4x10^{-6}$) (114). These results suggest that genetically and/or epigenetically dysregulated expression of *SREBF1* could be a relatively common feature of metabolic syndrome.

1.8.2 DNA Methylation in T2DM

Early studies of the epigenetics of T2DM typically employed a candidate gene approach. Such studies included analysing the insulin gene itself (*INS*) and other genes with known roles in insulin signalling, e.g. the glucagon-like peptide-1 receptor (*GLP1R*) (110). However, these approaches have now been largely replaced by genome-wide interrogations of CpG methylation.

There is overlap between CpG sites identified as significantly associated with T2DM and obesity. For example, the fat mass and obesity associated (*FTO*) gene, which is linked to obesity development, was identified as differentially methylated in the first large-scale EWASs of T2DM (115). Also, in a recent study where previously identified T2DM associated CpG sites were replicated in a novel cohort of 100 type 2 diabetics and 100 controls, only five out of 15 initially significant replicated sites remained significant after controlling for BMI (116).

A recent systematic review of EWASs of T2DM indicated that just over 100 CpG sites have been significantly associated with T2DM in studies of blood (116). EWASs of tissues directly relevant to T2DM, such as pancreatic tissue, have been less fruitful, yielding a total of 18 significant CpG sites to date. However, this is likely due to small sample cohorts (the largest one including a total of 35 individuals with T2DM) (116).

1.8.3 Histone Modifications and Non-Coding RNA in Obesity and T2DM

Further technological advancements are required before post-translational modifications (PTMs) to histones and ncRNA species associated with obesity and T2DM can be studied in a cost-effective way on a large scale. There is also a need for development of high-throughput methods of analysing cross-talk between epigenetic and genetic mechanisms, and between different types of epigenetic modifications.

Nevertheless, there are indications that histone modifications and their dysregulation are likely to be involved in T2DM. For example, several histone deacetylases called sirtuins (*SIRT1-7*) have been repeatedly identified as involved in insulin resistance, metabolic processes and inflammation (117, 118). One of the sirtuins, SIRT1, suppresses inflammation in both adipocytes and macrophages and improves glucose tolerance and reduces hyperinsulinaemia (119).

Recent research has also begun to disentangle the role of ncRNA species in obesity and T2DM. Case-control studies have demonstrated down- or upregulation of specific miRNAs in obesity and T2DM, some of which could emerge as useful biomarkers for disease progression (120). The roles of ncRNAs have also been demonstrated in studies on non-human mammals. For example, in a murine model, the brown fat long non-coding RNA (*Blnc1*) forms a core component of hepatic lipogenesis via the *LXR/SREBP1c* pathway, which is implicated in the development of NAFLD (121). In another murine model, obesity was associated with significantly elevated levels of the lncRNA Lnc-leptin, which is transcribed from an enhancer region upstream of leptin and which is required for leptin synthesis (122).

1.8.4 Conclusions from Epigenomic Profiling of Obesity and T2DM to date

The majority of large-scale EWASs of complex conditions so far have been performed on components of blood. As epigenetic signatures are highly tissue- and even cell-specific, the results of such analyses need to be interpreted with caution. In the context of obesity and T2DM, more relevant tissues to study are the liver, pancreas, hypothalamus and adipose tissue. Some studies have addressed this issue by attempting to replicate findings from EWASs performed in blood in more relevant metabolic tissues. One large epigenome-wide association study used the 450K human methylation array to identify 187 differentially methylated CpG sites associated with BMI in blood (106). They then found that methylation levels at these 187 loci correlated moderately to strongly between blood and metabolically relevant tissues, including included subcutaneous and omental fat, liver, muscle, spleen and pancreatic tissue (106). There was directional consistency in 120 out of the 187 sites for an association with BMI in both adipose tissue and blood ($p < 0.05$ after

Bonferroni correction). The authors suggest that the CpG sites identified in blood are therefore a reasonable reflection of their methylation levels in adipose tissue. Similarly, in liver, 114 of the 187 CpG sites showed directional consistency for association with BMI compared to blood (p = 0.001). It should be noted, however, that the overall correlation between methylation levels of the 187 CpG sites identified in blood compared to the other tissues ranged widely (R = 0.37− 0.93, p = 8.9×10^{-8} to 1.9×10^{-82}). Perhaps such inconsistency between tissues could serve as a cautionary note for future epigenetic investigations where blood is used as a surrogate for a tissue more relevant to the disease phenotype. Further, the fact that there was a correlation between findings in blood and other (more disease-relevant) tissues should have spurred investigation into whether the correlation could be explained by underlying genetic variation rather than BMI. Thus, there could be genetic variants, either at the at the CpG site itself or elsewhere in the probe sequence, that influenced CpG methylation at the sites of interest without being associated with BMI.

A significant challenge in interpreting results from EWASs is determining causality, i.e. whether a phenotype has occurred due to epigenetic alterations, or whether a phenotypic change has induced epigenetic change. Recent studies have begun to address this issue by using methods like Mendelian randomisation. The approach relies on the natural randomisation of genetic variants that occurs at conception and frequently uses SNPs identified through GWASs to infer causality as to whether particular CpG methylation signals identified in EWASs are likely to be a cause or effect of the phenotype of interest (123). This method was employed by Wahl *et al.* to investigate the potential causal relationships between DNA methylation in blood and BMI. A weighted genetic risk score was first calculated from a previously published GWAS of BMI, and then for each of the 187 identified significant CpGs, the effect of genetic risk score on methylation predicted through BMI was compared to the directly observed effect of genetic risk score on the CpG. An overall strong correlation between observed and predicted effects was seen ($R^2 = 0.81$, $p = 4.7 \times 10^{-44}$). This suggests that for the majority of the CpG sites, methylation in blood is the consequence rather than the cause of BMI variance (106).

A pertinent question in this context is what proportion of phenotypic variance could be attributable to epigenetic alterations. In a meta-analysis of 94 CpG sites identified as associated with BMI in a total of approximately 3000 individuals, it was estimated that the collective effect of methylation of these sites explained between 14-26% of the observed variation in BMI (105). Such estimates are likely to be refined as sample sizes increase and bioinformatic analyses improve. Still, these figures suggest that obesity is associated with substantial differences in DNA methylation.

In addition to those mentioned, there are several other caveats with results from EWASs of obesity and T2DM performed to date. For example, the Illumina HM450, which has been used in all large-scale obesity EWASs performed to date, only captures approximately 1.7% of the ~ 28 million CpG sites in the genome, and is heavily biased towards promoter regions. Future studies should use methylation arrays with improved coverage of enhancers and other regulatory regions, e.g. the MethylationEPIC array, until WGBS becomes a cost-effective alternative. Further, there should be increased interrogation into whether genetic variants could have influenced EWAS results, such as in studies of the correlation of epigenetic markers between tissues. Well-powered longitudinal analyses are likely to be key in this regard, and have the potential to detect biomarkers and epigenetic perturbations related to disease complications before the complications develop, allowing tailored interventions.

In summary, epigenome-wide association studies of components of the metabolic syndrome have largely focussed on CpG methylation in common conditions such as obesity and T2DM. The vast majority of studies thus far have been cross-sectional and thus unable to definitively determine causality. Studies have generally been performed on blood, and only a small proportion of CpG sites identified have been robustly replicated across cohorts. Despite these caveats, it seems that there are epigenetic variants associated with obesity or T2DM. Continued elucidation of these would improve our understanding of disease pathogenesis and identify biomarkers for disease progression.

1.9 Inheritance of Epigenetic Marks through Gametes

Paternal genetic factors influence fetal growth, but can paternal epigenetic marks acquired through the development of metabolic traits, be transmitted from father to offspring (67, 124)? This phenomenon is known as inter- and transgenerational epigenetic inheritance. Intergenerational effects occur when a fetus is exposed *in utero*, which can affect both the F1 and F2 generation (on the maternal side). On the paternal side, intergenerational effects refer to acquired or environmental changes to his germ cells that impact the F1 generation (125, 126). In strict terms, transgenerational effects refer to when epigenetic information is transmitted to the F2 generation (on the paternal line or in a maternal line in which exposure occurred only before conception) or F3 (on the maternal line when exposure occurs during pregnancy) generation and beyond (125). To convincingly show that germ cell alterations induce phenotypic changes in his offspring, several confounders need to be accounted for. These include changes in factors such as the composition of seminal fluid, the maternal reproductive tract at conception, the *in utero* environment and parental behavioural factors (127).

Transgenerational epigenetic inheritance in humans is controversial, in part due to the extensive epigenetic reprogramming that mammalian germ cells and preimplantation embryos undergo in order to generate an epigenetic state competent for totipotency (128). In other words, because a zygote requires the potential to differentiate into every tissue in the human body, its epigenetic makeup is reset between generations. Two distinct waves of epigenomic erasure occur; one shortly after fertilisation, during preimplantation development in the pronuclear zygote, and one in primordial germ cells (PGCs), during gonadal formation. Reprogramming of human primordial germ cells (hPGCs), including erasure of imprinting and epimutations, restores full germline totipotency (128) (Figure 1-13). DNA demethylation occurs both through passive loss of methylation during DNA replication, in the absence of methylating enzymes, and through active removal of methylation (129). Alongside demethylation, hPGCs undergo chromatin reorganisation and X reactivation (128). These processes mean that methylation levels in hPGCs are at basal level around week 5-7 of embryonic development (128).

In the peri-implantation period, the genome is re-methylated in a sex-, cell- and tissue-specific manner, allowing for lineage commitment, X-inactivation and establishment of imprinting. Recent research suggests that some loci escape the described genome-wide demethylation process, both in mouse and in human PGCs (128, 130). Notably, these 'escape loci' are preferentially found in genes controlling neurological and metabolic processes and are enriched for in gene bodies and regulatory regions (128). This implies a potential mechanism for environmentally acquired traits to be transmitted from one generation to the next.

In men, a third period of epigenetic reprogramming occurs with the onset of spermatogenesis and spans from puberty to adulthood (131). This may explain how changes in adult health and metabolism could impact on the sperm epigenome and gene expression in future generations.

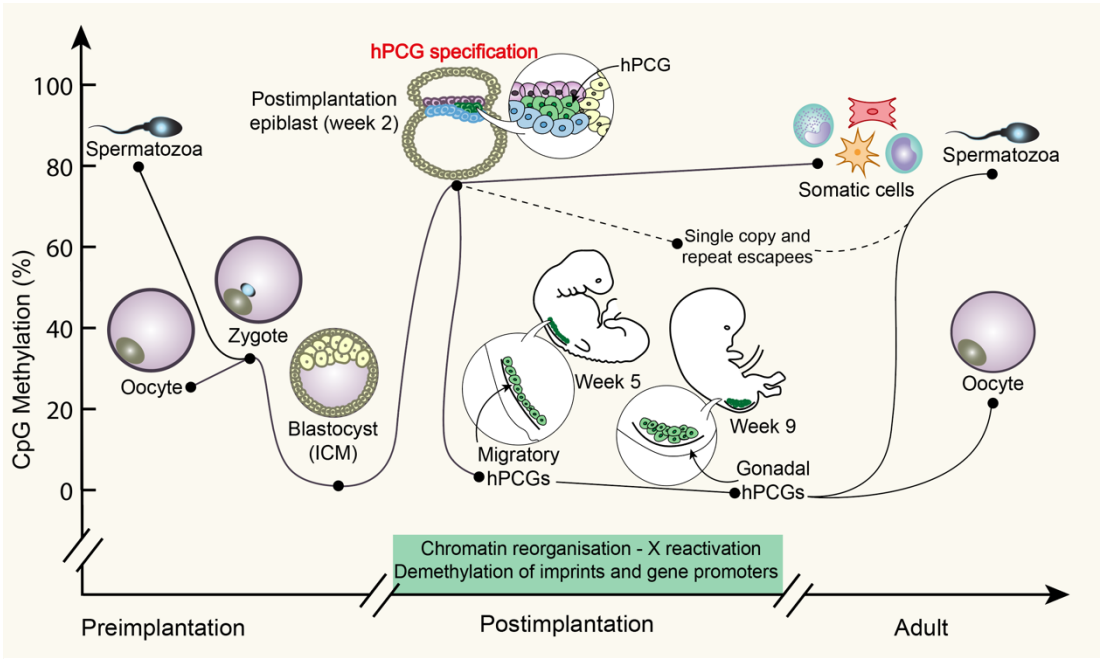


Figure 1-13. Schematic showing the dynamics of preimplantation and germline epigenetic reprogramming in humans.

hPGCs undergo the most comprehensive wave of global DNA demethylation shortly following fertilisation. A second wave of demethylation occurs in the primordial germ cells during gonadal development in the developing embryo. Adapted from (128).

1.10 Paternal Metabolic Syndrome, Epigenetic Inheritance and Fetal growth: Studies in Animals

Studies in non-human mammals have demonstrated that the sperm methylome can be altered by environmental and physiological change, including dietary alterations, toxins and even psychological stress (132-137). However, simply showing that an environmental factor is associated with germline epigenetic changes does not mean this change influences the next generation. Further, such associations are of little value in the context of intergenerational inheritance unless they also account for the widespread demethylation process that occurs between generations. More convincingly, other studies have suggested that acquired paternal traits may influence his offspring via alterations in his germ cells.

A seminal study by Ng *et al.* 2010 showed that a paternal high fat diet can induce lasting physiological changes in his offspring (138). Interestingly, only the female offspring of the HFD fed fathers were lighter at birth compared with offspring of fathers fed a control diet (CD). Furthermore, these F1 daughters had altered gene expression patterns in several hundred pancreatic islet cell genes and metabolic dysfunction at 6 and 12 weeks of age (138).

Further evidence that the effect of parental metabolic phenotype is gamete-driven comes from a clever in vitro fertilisation (IVF) study of genetically identical male and female mice fed combinations of a high fat diet (HFD), normal chow (NC) and low fat diet (LFD) as illustrated in Figure 1-14. Sperm and oocytes were isolated from F0 mice at the end of the dietary exposure and used for in vitro fertilization (IVF). The two-cell embryos obtained were transferred into healthy surrogate females to generate offspring (F1) from different parental combinations (127). The F1 pups were then all subjected to a high fat dietary challenge, and were followed up with regards to weight and metabolic health. In the F1 generation, female mice whose parents both had been subjected to an HFD were both heavier and more insulin resistant than those whose parents had been fed a CD (Figure 1-15). Indeed, they exhibited a blood glucose levels in response to an intraperitoneal glucose challenge and had a higher level of fat mass. Interestingly, F1 mice with only one parent subjected to a HFD

exhibited a similar, albeit tempered, metabolic dysfunction (139). These results show that the effect of diet induced metabolic dysfunction in both parents seems, at least in part, to be transmitted through diet induced changes in their gametes.

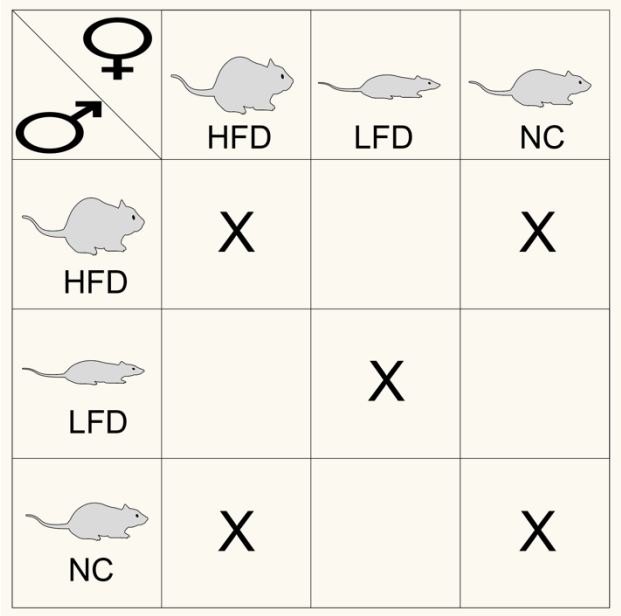


Figure 1-14. Schematic showing combinations of mating investigated in the study by Huypens *et al.*
Two cell embryos from the parental IVF combinations marked with an 'X' above were transferred to lean, healthy surrogate females to limit the impact of intrauterine effects on the offspring phenotype.
HFD = High Fat Diet, LFD = Low Fat Diet, NC = Normal Chow. Data derived from (127)

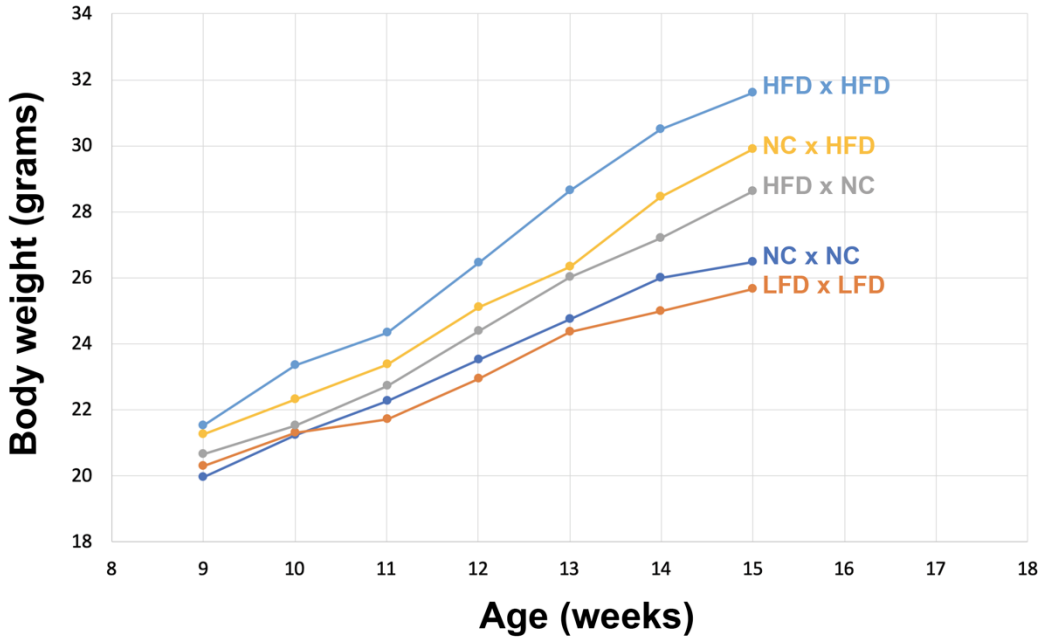


Figure 1-15. Body weight trajectories of female F1 mice during an HFD challenge between 9 and 15 weeks of age.

HFD = High Fat Diet, LFD = Low Fat Diet, NC = Normal Chow. Data derived from (127).

In another experiment, prediabetes was induced in male mice by feeding them a HFD and injecting them with a low-dose streptozocin before mating them with female mice fed a control diet CD (140). The offspring exhibited glucose intolerance and insulin resistance, and showed altered expression patterns of several hundred genes in pancreatic islet cells (140). Analysis of pancreatic islet cells identified changes in cytosine methylation at several insulin signalling genes, and these changes correlated with their expression patterns (140). Notably, analysis of paternal gametes showed an extensive overlap in the regions that were differentially methylated in offspring pancreatic islets and in sperm. It is possible that the phenotypic changes seen in offspring could be transmitted by other means, such as changes in seminal fluid or differences in mating behaviour. Nevertheless, this study supports the concept that it is possible for an environmentally-induced epigenetic signatures to be inherited through the paternal germline and alter offspring phenotype.

Martinez et al similarly identified an epigenetic signature on paternal gametes that was also present in metabolically important tissues in the offspring. In this study pregnant (F_0) mice were fed a half-normal caloric intake in late gestation, which resulted in growth-restricted offspring (IUGR-F_1) (141). The IUGR- F_1 mice developed obesity, mild hyperglycaemia and glucose intolerance in adulthood; an intergenerational effect. When the IUGR- F_1 males were mated with females fed a control diet, their resultant offspring (IUGR-F_2) also exhibited features of metabolic syndrome, including hyperglycaemia and hypercholesterolaemia. Further, expression profiles of IUGR-F_2 livers revealed 172 differentially expressed genes compared to control F_2 mice, several of which are implicated in lipid biosynthesis. Analysis of upstream transcription factors regulating these lipogenic genes showed that the liver X receptor alpha (*LXRA*) was down-regulated in the livers of IUGR-F_2 mice as a result of hypomethylation of 5' regulatory regions of the gene. Notably, the researchers then found this specific epigenetic signature in the spermatozoa of IUGR-F_1 males as well as in IUGR-F_2 fetal livers. In adult IUGR-F_2 mice, the signature was also present in white adipose tissue and pancreatic islet cells, but absent in skeletal cells, suggesting tissue specific postzygotic epigenetic reprogramming. Functional analyses of Lxra using knockout models supported that decreased levels of Lxra has

downstream effects such as those observed in adult IUGR-F_2 mice, including moderate hypertriglyceridemia and increased hepatic free cholesterol levels (141).

In summary, there are at least two studies in mammals where distinct diet-induced epigenetic signatures are found both in paternal gametes and in metabolically relevant tissues in his offspring. The lack of concordance of results in these studies may reflect differences in timing and type of dietary insult to the parent, different strains of mice or different methods of epigenomic investigation. However, these findings also raise several questions. How did the epigenetic mark seen in paternal gametes either escape the extensive demethylation process that follows fertilisation, or become re-established post-fertilisation? And if the mark does escape this process, then why is it not present in all offspring tissues?

It should also be said that few of the studies discussed above have been convincingly replicated, and that some researchers suggest that stochastic epivariation rather than generational effects of diet may underlie the observed phenotypic effects (142, 143). Further, the role of DNA methylation in the context of intergenerational epigenetic inheritance has been challenged by recent research indicating that small RNA species may be more instrumental in generational effects (144, 145).

1.11 Paternal Metabolic Syndrome, Epigenetic Inheritance and Fetal growth: Studies in Humans

The evidence for epigenetic inheritance through gametes in humans is sparse at best. However, several studies point towards the possibility that the sperm epigenome is dynamically remodelled during an individual's lifetime. One study found that spermatozoa of obese men carry a distinct epigenetic signature compared to that of lean men (146). Differences principally affected genes involved in the central control of appetite. Following extreme weight loss by bariatric surgery, this epigenetic profile was altered (147). However, this study can be criticised for only including 5 men without a control group. Furthermore, none of the epigenetic changes following weight loss met the threshold for epigenome-wide significance after correcting for multiple testing. Two longitudinal studies have indicated that 3 months of aerobic

exercise could be associated with remodelling of the human sperm methylome (148, 149). These studies were small-scale and there was little overlap between their results. Nevertheless, they point towards the potential for dynamic gamete remodelling in response to metabolic alterations in a tissue that has the potential to influence the next generation.

If epigenetic change in gametes underlie the association between paternal obesity and T2DM, and in-utero growth of his offspring, then infants born to obese or diabetic fathers would be expected to carry epigenetic profiles that are distinct from those of appropriately grown offspring. In support of this hypothesis, infants affected by fetal growth restriction (FGR), are more likely to have insulin-resistant fathers, and have a distinct DNA methylation profile in umbilical cord blood relative to appropriately grown offspring (150). Further, in a cohort of 1,046 infants from the Norwegian Mother and Child Cohort Study (MoBa), there was differential methylation of 19 CpGs in cord blood DNA associated with birth weight (151). Differential methylation of 2 CpGs on *ARID5B* (AT-rich interactive domain 5B), a gene associated with low birth weight and decreased adipose tissue in mice, were inversely correlated with birth weight (151).

Offspring born to parents exposed to caloric restriction during the Dutch Hunger Winter (see section 1.2), had 181 differentially methylated regions (DMRs) in an epi-genome-wide DNA methylation analysis of their whole blood compared with their same-sex siblings not conceived during the famine (152). Gene ontology analyses of these regions revealed enrichment for pathways involved in positive regulation of growth and lipid and cholesterol metabolism, consistent with the more atherogenic lipid profile in these individuals (152). Whether epigenetic differences in offspring affected by the Dutch Hunger Winter are a consequence or cause of altered metabolic phenotype is unclear.

In summary, there is only circumstantial and not conclusive evidence of intergenerational inheritance of acquired paternal epigenetic changes influencing

fetal growth and birth weight of his offspring. Prospective studies in humans are lacking, but hard to conduct.

1.12 The Human Sperm Epigenome

The human sperm epigenome is less well characterised than more readily available somatic tissues, such as blood. As human semen is only routinely analysed in reproductive medicine settings, most research has focused on the human sperm methylome in relation to fertility. There has been a general lack of robust, reproducible findings across studies. Sperm represents the most relevant tissue to study when investigating the potential for acquired paternal traits to impact the next generation.

Blood is the most commonly studied surrogate tissue when sperm is not available. For example, Soubry *et al.* investigated the association between paternal obesity and methylation status of seven imprinted genes in offspring cord blood (153). They showed that periconceptual paternal obesity was significantly associated with altered methylation status of three of the genes studied (153). However, as the study analysed paternal blood rather than germ cells, it is difficult to draw conclusions regarding the mechanism behind the identified epigenetic effects. In spite of this, the authors suggest that the results likely indicate that paternal obesity alters the DNA methylation profile of his germ cells such that it may impact the phenotype of the next generation:

"… The significant and independent association between paternal obesity and the offspring's methylation status suggests the susceptibility of the developing sperm for environmental insults. The acquired imprint instability may be carried onto the next generation and increase the risk for chronic diseases in adulthood." (153)

Such conclusions are tentative owing partly to the highly tissue-specific nature of epigenetic signatures, and partly to the widespread demethylation process that occurs between generations (section 1.9). Future research would thus be aided by either investigating paternal germ cells directly, or, as an intermediate approach, to better characterise the epigenetic correlation between surrogate tissues such as blood and spermatozoa.

1.12.1 Candidate Gene Analyses of DNA Methylation in Human Spermatozoa

Early studies of the human sperm methylome followed epidemiological reports of an association between assisted reproductive technologies (ART) and an increased risk of imprinting disorders in the offspring (154). It was hypothesised that spermatozoa from males with abnormal semen parameters who conceived with the aid of ART could harbour a greater frequency of abnormally imprinted genes that would impact on the phenotype of the offspring. In this regard, several candidate gene studies were directed towards the imprinted gene clusters. These included 15q11-q13, implicated in the Prader-Willi and Angelman syndromes, and the *H19/IGF2* locus, which is implicated in the Beckwith-Wiedemann and Silver-Russell syndromes (155-159). Other candidate gene analyses focused on genes involved in spermatogenesis, early embryogenesis and DNA methylation (160-162). Whilst some of these studies suggested that imprinting disorders were more frequent in sperm from subfertile men, other studies found no evidence between imprinting disorders and the outcome of ART (157, 158).

The cost-effectiveness of candidate gene studies has made them an attractive approach to studying small numbers of genes thought to be involved in a particular biological process. However, these hypothesis-driven approaches suffer from several methodological limitations. These include publication bias (as negative results from a small-scale study are less likely to be published), incomplete understanding of the biological pathways and lack of power to detect modest effects.

1.12.2 Genome-wide Analyses of DNA Methylation in Human Spermatozoa

Epigenome-wide array-based methods have interrogated CpG methylation across the spermatozoal methylome, typically in case-control studies of fertile versus infertile/subfertile males (163-167). As the majority of participants in these studies have been recruited from reproductive medicine settings, the results may not be broadly relevant.

It is unlikely that studies similar in magnitude to EWASs performed on blood in the context of obesity and T2DM will ever be performed on spermatozoa. An

intermediary step in understanding how metabolic syndrome in humans could affect the epigenetic make-up of sperm would be to perform a detailed characterisation of the epigenetic covariation between the two tissues. This would allow findings obtained from analyses in blood to be compared with the relevant CpG sites in sperm. For example, if paternal obesity is robustly associated with the methylation status of an individual CpG site in both blood and sperm, intergenerational effects of acquired paternal obesity become more likely, compared with a lack of tissue correlation.

The largest study to characterise DNA methylation co-variation in matched human sperm and blood components (B cells) included just eight males (168). This study demonstrated large-scale differences between the DNA methylome of sperm and blood, such that spermatozoa displayed methylation levels towards the extremes, i.e. methylation levels below 20% or above 80%. This observation already questions whether analyses of blood can be used to infer DNA methylation signatures of germ cells. The analysis further convincingly contradicted some earlier candidate gene studies that had, for example, indicated that abnormal spermatozoa exhibit epigenetic dysregulation of the Deleted in Azoospermia-Like (*DAZL*) gene, and that abnormal spermatozoa frequently exhibit imprinting abnormalities (161, 169). Thus, in spite of its small sample size, this study provided a valuable contribution to the field, and highlighted the need for larger-scale, genome-wide analyses of germ cells and matched somatic cells in future research.

1.13 Rationale for my PhD Project

Small for gestational age infants are at increased risk of developing metabolic disorders, including T2DM and cardiovascular disease, in adulthood (52, 65, 170). Epidemiological studies suggest that paternal obesity and/or insulin resistance predisposes his offspring both to poor growth *in utero* and future metabolic disorders (68, 69). Genetic association studies have identified an overlap between genetic variants that predispose to birth weight and risk of T2DM (171). However, SNPs only explain a small proportion of the variance of these traits, suggesting that gene-environment, or epigenetic, factors, may be influential in the development of the disease phenotype.

Large-scale epigenome-wide association studies have identified CpG sites that are differentially methylated between lean and obese individuals (104-106). However, most studies have been performed using blood samples, which are of questionable importance in the context of intergenerational inheritance. Further, most identified CpG sites to date remain unreplicated across study cohorts.

Animal studies suggest that the association between paternal obesity and T2DM and fetal growth could be mediated through epigenetic changes to spermatozoa that are passed from a father to his offspring and affect metabolism in the developing fetus (140, 141). Such studies are lacking in humans. Indeed, epigenetic studies of human spermatozoa have largely been focussed around fertility, and the knowledge of how acquired traits could influence the human spermatozoal methylome is patchy. A systematic review and summary of studies of DNA methylation in human spermatozoa would provide a useful resource for understanding how the human spermatozoal methylome could be influenced by acquired traits and their potential to affect the next generation.

Studies of the human sperm methylome thus far have frequently been hampered by biased approaches and small sample sizes. A comprehensive, unbiased characterisation of the human sperm methylome in healthy, fertile males would provide novel insights into the epigenetic profile of the tissue that is most relevant in understanding generational effects of acquired paternal traits. Further, a comparison between the human sperm methylome and that of matched blood samples would improve our understanding of the circumstances under which surrogate tissues such as blood could be used for studies of intergenerational inheritance. Also, to analyse whether the covariation between DNA methylation in sperm and blood is influenced by obesity would yield insights into the potential for acquired paternal traits to influence his germ cells.

Overall, an improved understanding of the association between paternal metabolic syndrome and his offspring's growth *in utero*, and the mechanisms behind this

association, has the potential to improve the primary prevention of obesity and T2DM, which are major global public health concerns.

1.14 Hypotheses

1) Paternal obesity and/or insulin resistance increases the risk of fathering small for gestational age (SGA) offspring

2) A. The human sperm epigenome displays large-scale overall differences compared with matched somatic tissues such as blood
   B. At a select number of CpG sites, there is a high DNA methylation correlation between the tissues

3) Obesity influences the sperm DNA methylome

4) Obesity is associated with a robust, reproducible DNA methylation profile in blood

1.15 Thesis Aims

1) To identify paternal metabolic risk factors for low birth weight offspring in a prospective cohort study of 500 mother-father-offspring trios

2) To systematically review and summarise current knowledge of the human sperm methylome and its potential for being influenced by acquired traits

3) To perform the largest to date characterisation of the DNA methylome in matched samples of human sperm and blood

4) To identify specific CpG sites that show a high degree of DNA methylation correlation between blood and sperm in order to inform future research of potential intergenerational effects where sperm may not be available

5) To undertake a case-control study of DNA methylation covariation in sperm and blood between lean and obese males in order to investigate whether obesity is associated with an altered DNA methylation profile in male germ cells

6) To generate a robust, replicable obesity associated DNA methylation profile in blood by replicating CpG sites previously identified as significantly associated with obesity in a novel cohort of 96 lean and 96 obese males

Chapter 2
Research Methods and Materials

2.1 Introduction

In order to test the hypotheses stated in section 1.14, I carried out four separate but linked investigations.

First, I performed a prospective cohort study, hereafter referred to as 'The Dad's Health Study', where I recruited couples due to have a child at UCLH in order to investigate the association between paternal metabolic disease and offspring birth weight. The parents were phenotyped with regards to metabolic health, and the pregnancies were followed up with regards to pregnancy outcome and offspring weight at birth. This study is described in sections 2.3-2.8.

Second, I performed a systematic review of studies investigating DNA methylation in human spermatozoa following established guidelines. This allowed me to summarise current knowledge, generate recommendations for future research, and inform the next stages of my project. The methods of performing the systematic review are discussed in section 2.9.

Third, I investigated whether any observed association between paternal metabolic health and offspring birth weight could be mediated by spermatozoal DNA methylation changes passed from the father to his offspring. With this aim, I performed an extensive characterisation of the spermatozoal DNA methylome and compared it to that of matched blood samples. I characterised this sperm-blood methylation covariation both in lean, fertile males and in overweight/obese, fertile males. I then compared the sperm-blood DNA methylation covariation profiles of lean and obese males. This study is described in sections 2.10-2.12.

Lastly, I aimed to establish whether obesity is associated with a robust, replicable DNA methylation signature in blood. With this aim, I compiled results from previous studies investigating genome-wide CpG methylation in obesity and validated these results using a multiplex bisulfite-PCR-sequencing microfluidics based assay in peripheral blood from a novel cohort of lean and obese males. This study is described in sections 2.13-2.15.

2.2 Declarations

Recruitment and sample collection for the Dad's Health Study was performed by myself and research midwife Anna Greco. Sample processing and DNA extraction was also performed by myself and research midwife Anna Greco. Statistical analyses were carried out by myself under guidance from Jim Tyson, Senior IT Trainer at UCL, and Professor Aviva Petrie at the UCL Eastman Dental Institute.

For the systematic review of studies of DNA methylation in human sperm, the search, filtering of results and summarising of studies was first performed by myself. The search was then repeated independently by Dr Amy Danson, Queen Mary University London (QMUL). Dr Amy Danson also added to the grading of studies and we together discussed overall conclusions and recommendations for future research. Dr Sarah Marzi, QMUL, had an advisory role in the search process and evaluation of evidence.

For the study of DNA methylation covariation of sperm and blood, I collected and processed the sperm and blood samples together with research midwife Anna Greco. Bisulfite conversion of DNA from blood and sperm samples was performed by myself and Ama Brew, research technician at the Blizard Institute, QMUL. The Illumina MethylationEPIC array was processed by Yasmin Panchbhaya at UCL Genomics, Great Ormond Street Institute of Child Health. Methylation data preprocessing and analysis was performed by Dr Sarah Marzi at the Blizard Institute, QMUL.

For the analysis of obesity associated CpG sites in blood, I and research midwife Anna Greco collected the blood samples and phenotyped participants. Blood samples were also collected by Dr Donna Santillan at the Women's Health Tissue Repository, University of Iowa Health Care. DNA extraction from all blood samples was performed by myself and Research Midwife Anna Greco. Primers for the multiplex bisulfite PCR sequencing microfluidics-based assay were identified from previous research by Dr Michelle Holland and Dr Sarah Marzi at QMUL under guidance of Professor Vardhman Rakyan. Primers were tested for efficacy and specificity by Adrian Signell at Kings College London under supervision from Dr Michelle Holland. Bisulfite conversion of blood samples was performed by Theodoros Xenakis at QMUL.

The multiplex bisulfite-PCR-sequencing microfluidics based assay, library preparation and subsequent Next Generation Sequencing (NGS) was performed at the Genome Centre Facility at QMUL. Methylation data preprocessing and analysis was performed by Dr Sarah Marzi at the Blizard Institute, QMUL.

2.3 The Dad's Health Study

A prospective cohort study known as 'the Dad's Health Study', detailed in sections 2.3-2.8, was performed to investigate the association between paternal metabolic health and offspring birth weight. Parental and offspring phenotypic measurements were recorded as described in sections 2.6 and 2.7.

DNA was extracted from buffy coat from parental blood samples and offspring cord blood samples as described in section 2.8. Paternal semen samples were collected and analysed as described in section 2.9. DNA methylation was analysed using the Illumina MethylationEPIC array as described in 2.11.

2.3.1 Study Design

A prospective cohort study, The Dad's Health Study, was undertaken of lean, overweight and obese fathers and their female partners (the mothers) to investigate the association between paternal metabolic health and offspring birthweight. The study was carried out at University College London Hospital between May 2016 and May 2019. Favourable ethical approval for the study was granted from the South East Coast - Surrey Research Ethics Committee on 28 September 2015 (REC reference number 15/LO/1437, IRAS project ID 164459). The study was also registered with the UCLH Joint Research Office (Project ID 15/0548). All participants provided written, informed consent.

2.3.2 Pilot Study

A pilot study of matched blood and semen samples from eight healthy male volunteers was performed in October-December 2015, prior to commencing recruitment for The Dad's Health Study. The pilot study sought to confirm that the techniques for processing of and DNA extraction from blood and semen, detailed in

sections 2.8 and 2.9 were adequate to 1) yield purified samples of human semen such that only the motile portion of sperm was used for DNA extraction while contaminating cells and seminal fluid was discarded and 2) yield sufficient quantities of high quality DNA from blood and semen for downstream epigenetic analyses. Recruitment for The Dad's Health Study commenced following successful completion of the pilot study.

2.3.3 Statistical Methods

Statistical analyses were carried out using RStudio version 1.1.456. Results were independently verified by Dr Aviva Petrie at the UCL Eastman Dental Institute Biostatistics Unit using STATA 15 (StataCorp LLC, Texas USA).

Lean, overweight and obese male participants were assessed for comparability with simple descriptive statistics. Baseline characteristics of the three groups of male participants and their female partners (the mothers) were summarised with means and 95% confidence intervals. For categorical variables, frequency counts and percentages were given. When comparing the means of two variables an unpaired t-test was used for continuous variables and a 2x2 table for categorical data. Two tailed P-values <0.05 were considered statistically significant.

Multivariable logistic regression was used to analyse the impact of paternal metabolic profile on offspring birth weight. Prior to study recruitment, the primary paternal variable hypothesised to be associated with offspring birth weight was BMI. Secondary variables to be explored were other components of the metabolic syndrome, including paternal insulin resistance measured using HOMA (detailed in section 2.7), waist circumference, blood lipid levels, blood pressure as well as paternal smoking (12, 69).

2.3.4 Sample Size Calculations

Sample size calculations were undertaken using *'Sample Size Tables for Clinical Studies, 3rd Edition'* (172). Paternal obesity was hypothesised to either be associated with an overall reduction in mean offspring birth weight, or with an increase in the

proportion of offspring born small for gestational age (SGA). Thus, two different approaches to determining an adequate sample size were employed.

In the first scenario, assuming that the smallest difference in birth weight that is clinically significant is a reduction of 150 g, obese paternity would be associated with an overall mean offspring birth weight of 3350 g compared with 3.5 kg for offspring born to lean fathers (standard deviation 400 g) (173). If using a two-sample t-test to compare the means of offspring birth weights, this would require the recruitment of 151 obese and 151 lean fathers, i.e. a total sample size of 302, to achieve 90% power at 5% significance.

Alternatively, paternal obesity could be associated with an increased proportion of offspring born SGA. Previous research has indicated that paternal obesity (BMI $\geq$ 30 kg/m^2) may be associated with a 24.8% (50/321) risk of fathering small for gestational age offspring, compared to a 9.8% (159/1631) risk of fathering a small for gestational age offspring among lean men (BMI <25 kg/m^2) (68). Assuming that the smallest difference in the proportion of offspring born SGA between the two groups of fathers (obese versus lean) that is clinically significant is 10%, and assuming that paternal obesity would be associated with a 20% risk of fathering offspring below the 10th customised centile whereas 10% of offspring born to lean fathers would be born SGA, 219 fathers in each group (438 in total) would be required to detect a 5% difference with 80% power.

A proportion of those recruited would not be included in analyses of the impact of paternal metabolic health on offspring birth weight for reasons such as miscarriage, multiple pregnancy that was unknown at the time of recruitment and transfer of care to other hospitals. Allowing for these and other causes of participant drop-out, a total recruitment target of 500 male participants (250 lean versus 250 obese) was felt to be justified.

In an audit performed at UCLH antenatal clinics in 2015 prior to recruitment, it was identified that a significant group of otherwise eligible participants were of an

intermediate (overweight) BMI of 25-30 kg/m^2. In order to investigate whether there was a linear, reverse correlation between paternal BMI and offspring birthweight as well as achieve a sample population that was more reflective of the male population as a whole, it was decided to include overweight men as a separate group in addition to those with a lean and obese BMI.

2.4 Feasibility of Recruitment

An audit performed in UCLH antenatal clinics in 2015 showed that approximately 40% of fathers were lean (BMI 18-25 kg/m^2), 40% of fathers were overweight (BMI 25-30 kg/m^2) and 20% of fathers were obese (BMI $\geq$ 30 kg/m^2). Around 6500 babies are delivered at UCLH each year (174). Assuming that 1 in 10 of obese fathers (130 per year) and 1 in 20 lean fathers (130 per year) would take part in the study, a two-year recruitment period was felt to be justified in order to recruit approximately 500 male participants and their female partners in total.

2.5 Recruitment Criteria

In order to minimise significant confounding factors that would be expected to impact on fetal growth independent of paternal metabolic profile, only fathers who met the following criteria were included in The Dad's Health Study:

- Aged 18 to 50 years
- No significant medical problems
- No drug, alcohol or substance abuse

The following inclusion criteria were applied to the female partners (the mothers):

- Aged 18 to 45 years
- No significant medical problems
- No drug, alcohol or substance abuse

The following inclusion criteria were applied to the pregnancy:

- Singleton pregnancy
- Natural conception

Men whose female partners who were due to deliver a baby at UCLH were approached in the antenatal clinics or responded to a research poster. The majority of participants were recruited around the time of their female partners first ultrasound scan (10-14 weeks of gestation).

Participating fathers were offered a 'Well Man Health Check', which I and Research Midwife Anna Greco performed, to measure phenotypic variables associated with metabolic health.

2.5.1 Exclusion Criteria

Offspring whose growth was likely to be affected by factors likely to confound any impact of paternal metabolic profile on offspring birth weight were excluded. Thus, the following exclusion criteria were applied:

- Multiple pregnancy
- Known maternal drug or excess alcohol intake
- Pregnancies where the mother is affected by a significant medical condition, such as cardiac, endocrine, renal or haematological disease
- Know fetal congenital infections such as CMV or toxoplasmosis
- Known congenital chromosomal abnormalities
- Known congenital structural malformations
- Pregnancies conceived via artificial reproductive therapies (ART)
- Maternal gestational diabetes

2.5.2 Diagnosis of Fetal Growth Restriction

Fetal Growth Restriction (FGR) is defined as failure of a fetus to meet its growth potential in the womb, and is associated with both long- and short term adverse health outcomes (175). A small for gestational age (SGA) infant is commonly defined as one whose weight at birth is less than the 10^{th} centile for gestational age. It

remains a challenge to distinguish between infants who are constitutionally small and those affected by true growth restriction (175).

In order to improve our identification of infants affected by FGR rather than being constitutionally small, as well as minimise the impact of maternal factors that could confound the effect of paternal metabolic profile on offspring birth weight, customised birth weight centiles were used instead of raw birth weights (176). Available from the Perinatal Institute, these customised centiles take into account maternal height, weight, ethnicity, offspring sex and gestational age, giving a more accurate representation of the ability of an infant to have met its intra-uterine growth potential (177). In the present study, infants were classified as SGA if their customised birth weight centile was <10, classified as AGA if their customised birth weight centile 10-90 and classified as LGA if their customised birth weight centile was ≥ 90.

In addition, mothers who developed gestational diabetes mellitus (GDM) during their pregnancy were excluded from the study, as this condition is associated with significant effects on fetal growth (section 1.3). Thus, in line with UCLH criteria for diagnosing GDM, participating mothers who had a 26-28 week glucose challenge test result that exceeded 7.8 mmol glucose per litre followed by a confirmatory glucose tolerance test (GTT) were excluded (178).

2.6 Study Protocol

The general outline of study procedures is illustrated in Figure 2-1. All phenotypic measurements of male participants (fathers) were performed in UCLH antenatal clinics using the same or identical equipment. The majority of men were studied around the time of their partner's first ultrasound scan (10-14 weeks into gestation). As fasting measurements of blood glucose and insulin were required, male participants were asked to fast overnight or for at least 8 hours prior to the study. Each study visit took approximately 30 minutes to complete. All participants were given a Participant Information Sheet and provided written, informed consent prior to commencing the investigations.

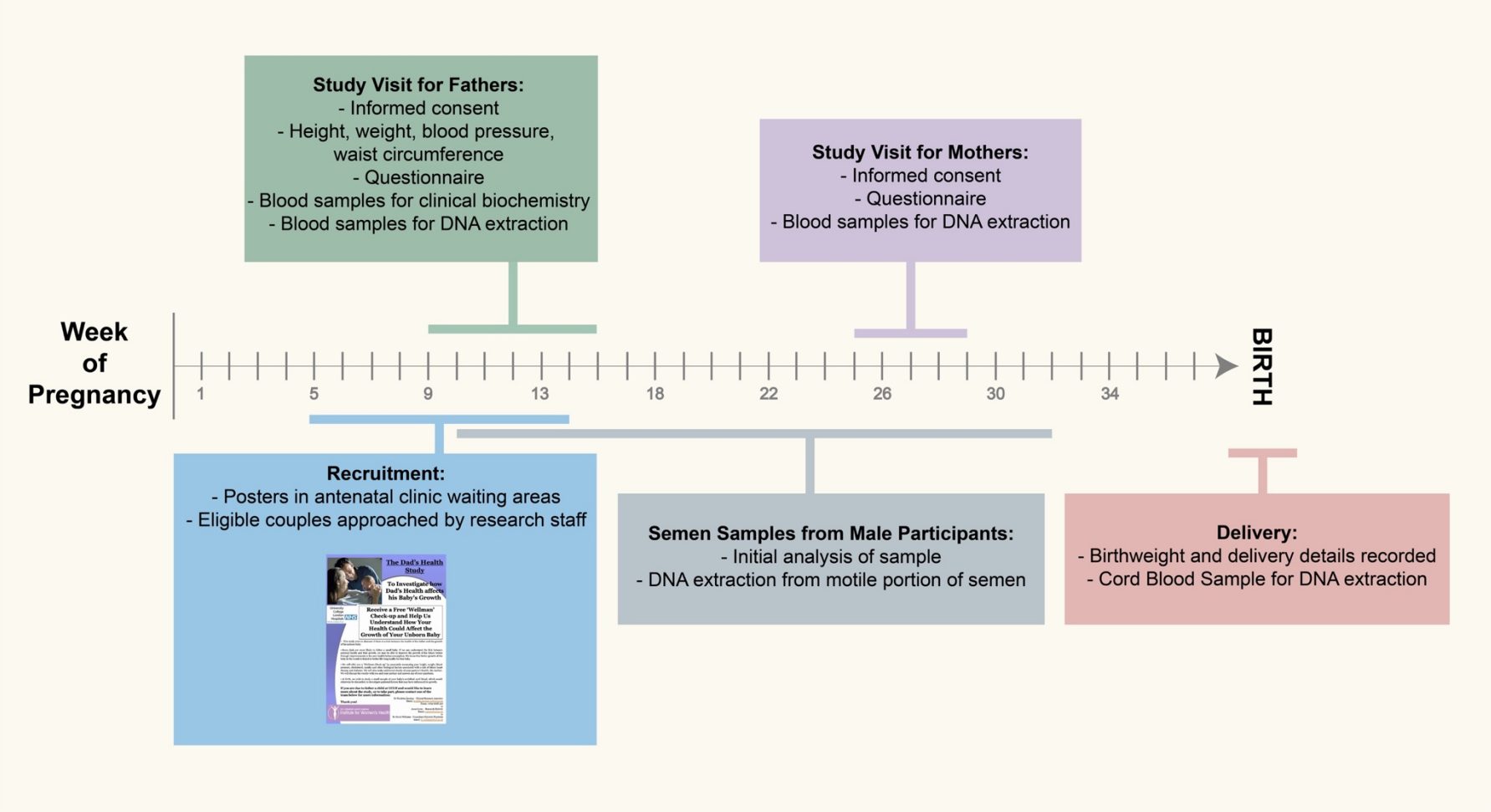


Figure 2-1. Overview of recruitment, timing of study visits and phenotype data collected for The Dad's Health Study.

Male participants completed a questionnaire enquiring about past medical, family and treatment history (Appendix 1). Study participants' own birth weight was recorded as remembered personally or from a parent. Height and weight were measured and recorded. Waist circumference was measured between the lower margin of the lowest palpable rib and the top of the iliac crest, using a stretch-resistant tape according to WHO guidelines (179). After resting, two measurements of blood pressure (BP) were taken 15 minutes apart.

Six vacutainers (20-25 mL blood in total) of whole blood were taken from participating males (fathers). Three (one purple top EDTA-coated, one gold top serum-separator tube (SST) and one grey top fluoride preservative) vacutainers were sent to the UCLH clinical biochemistry laboratory within one hour of venepuncture for fasting measurements of insulin, glucose, haemoglobin, c-reactive protein (CRP), haemoglobin A1c (HbA1c), lipid levels, liver function and kidney function. Insulin resistance was calculated using HOMA as detailed in section 2.7.

Feedback to participants about results was available. Clinically significant results were conveyed to the participants as a priority along with relevant support literature and advice. Where appropriate and with the participant's consent, general practitioners (GPs) were contacted on behalf of the participants.

Two purple and one gold top vacutainers were ultracentrifuged for 15 minutes at 1000g within one hour of venepuncture. Centrifugation of the purple top (EDTA-coated) vacutainers produced three distinguishable layers; the uppermost, clear layer consisting of plasma, the intermediate layer consisting of buffy coat and the lowermost layer consisting of concentrated erythrocytes (Figure 2-2). Buffy coat is a leukocyte-enriched fraction of whole blood and was used for DNA extraction The Dad's Health Study in preference to whole blood as this enhances DNA yield by approximately 5-10 times (180).

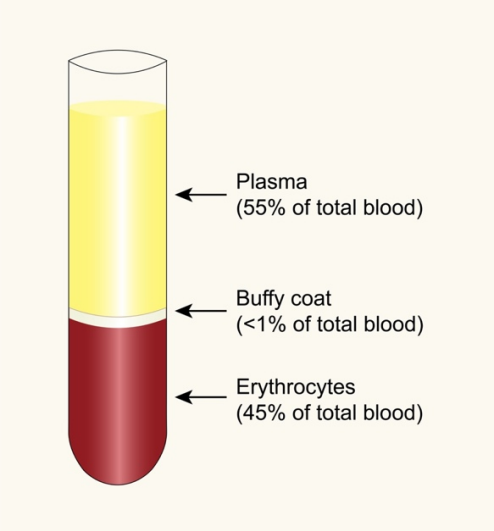


Figure 2-2. Separation of components of whole blood following ultracentrifugation of EDTA-coated vacutainers.
Whole blood separates into three distinct layers when ultracentrifuged for 15 minutes at 1000g; plasma, buffy coat and erythrocytes. Buffy coat was used for DNA extraction in The Dad's Health Study.

For each participant, two aliquots of plasma were obtained from the EDTA-coated vacutainers and stored in −80°C. Two aliquots of buffy coat were also obtained from the EDTA-coated vacutainers, one of which was used for DNA extraction (section 2.8.1). The aliquots of buffy coat not used for DNA extraction as well as the extracted DNA were thenceforth also stored in −80°C. Two aliquots of serum were obtained from the gold top SST vacutainer following ultracentrifugation and also stored in −80°C.

All female participants (mothers) were given a Participant Information Sheet and provided written, informed consent. The majority of female participants were invited for a study visit in conjunction with their 26 to 28 week GCT. Inviting participating females for a study visit around this time ensured that they were seen at a similar time in pregnancy, did not require an additional hospital appointment and were recruited well before their estimated delivery date (EDD), facilitating planning for later cord blood collection by placing a sticker and a leaflet in their antenatal notes. Some participating females instead opted to be seen around the same time as their male partners came to UCLH for the study visit (10-14 weeks into gestation), or in conjunction with another antenatal appointment. Female participants completed a

questionnaire enquiring about obstetric, medical, family and treatment history (Appendix 2).

A total of three vacutainers (two purple and one gold top) were obtained from participating females. These were ultracentrifuged within one hour of venepuncture and samples of buffy coat, plasma and serum were obtained and stored as described for male participants. DNA was extracted from buffy coat as described in section 2.8.1. Clinical biochemistry information required for participating females was obtained from electronic antenatal records, and included measurements of serum glucose at the initial antenatal visit (approximately 5-9 weeks of gestation), and results from the GCT or GTT. Information on ethnicity for both male and female participants as well as maternal height and weight, obstetric history and estimated date of delivery was obtained from electronic antenatal records.

Cord blood samples were obtained shortly following the delivery of the placenta, either by myself or by the attending midwife. A 10 mL sample was obtained from the umbilical arteries or vein using a syringe. Samples were ultracentrifuged for 15 minutes at 1000 g, aliquoted, processed and stored as described for parental samples. If any delay was expected between obtaining the cord blood sample and DNA extraction, samples were stored at 4° C.

2.7 Insulin Resistance and Type 2 Diabetes Mellitus

Insulin resistance (IR) is as a precursor to the development of Type 2 Diabetes Mellitus (T2DM), as well as forming part of the diagnostic criteria for the Metabolic Syndrome as discussed in section 1.1.

2.7.1 Measuring Insulin Resistance

The gold standard method for assessing insulin resistance is the hyperinsulinemic euglycemic glucose clamp. In brief, this involves intravenous administration of insulin at a constant, hyperinsulinemic rate and intravenous administration of glucose at variable rates to acquire and maintain a steady-state of euglycaemia. The rate of glucose infusion correlates to insulin sensitivity (181). Although a sensitive and

reproducible test for detecting IR, the hyperinsulinemic euglycemic glucose clamp is expensive and time- and labour intensive, and therefore inappropriate for routine clinical applications or larger-scale research studies (181). Thus, several indirect methods of assessing insulin resistance, which require only single measurements of fasting glucose and insulin, have been developed. Among these, the Homeostatic Model Assessment (HOMA) is among the most widely used and extensively validated (182).

2.7.2 Homeostasis Model Assessment (HOMA) Measurements

Homeostatic model assessment (HOMA) is a method for assessing insulin resistance (IR) and β cell function from basal (fasting) glucose and insulin concentrations. Originally described in 1985, HOMA values are derived from a mathematical assessment of the interaction between β cell function and IR in an idealised model that is then used to impute steady-state insulin and glucose concentrations. Using this calculated relationship between β cell function and IR, the model can then be used to estimate β cell function and IR from any given values of fasting insulin and glucose measurements (182).

The original model (HOMA1) used a simple equation for estimating insulin resistance from fasting blood glucose and insulin concentrations:

$$HOMA - IR = \frac{glucose\ \left(\frac{mmol}{L}\right) x\ insulin\ \left(\frac{\mu U}{L}\right)}{22.5}$$

The denominator 22.5 is a normalising factor derived from the product of a 'normal' fasting plasma glucose in a healthy individual (4.5 mmol/L) and a 'normal' fasting plasma insulin in a healthy individual (5 μU/L). The output is thus calibrated to give a normal IR of 1.

It should be noted that neither fasting insulin levels nor HOMA-IR values are normally distributed. Thus, log(HOMA-IR) is frequently used to transform the skewed

distribution of values to yield a more linear correlation with glucose clamp estimates of insulin sensitivity (181).

The model was updated from HOMA1 to HOMA2 in 1996 (182, 183). HOMA2 has nonlinear solutions and takes into account variations in hepatic and peripheral glucose resistance, increases in the insulin secretion curve for plasma glucose concentrations above 10 mmol/L and the contribution of circulation proinsulin, thus yielding a model closer aligned to physiological conditions (183, 184). Further, whereas the HOMA1 was calibrated based on an insulin assay used in the 1970's, HOMA2 has been recalibrated in line with current insulin assays (182). Therefore, whilst HOMA1 functions well to estimate relative change in IR when serial measurements are taken in the same individual, the corrected nonlinear model of HOMA2 is more accurate when assaying absolute insulin resistance or β cell function (183). HOMA2 values can be calculated using the calculator or excel plugin available from the Diabetes Trials Unit (184). Despite the update from HOMA1 to HOMA2, most recent large-scale trials have continued using the HOMA1 system for measuring insulin resistance, potentially due to the more straightforward calculation required and in order to make results more comparable to those in previous research. Therefore, both HOMA1 and HOMA2 values were calculated for participants included in the Dad's Health Study.

2.7.3 Validation of HOMA-IR

HOMA measurements have been extensively employed and validated in large scale epidemiological analyses as appropriate measurements of insulin resistance when taking single fasting measurements of blood glucose and insulin (183). HOMA-IR values show a strong correlation with glucose clamp estimates of insulin sensitivity, especially when log transformed (181). In a recent meta-analysis of surrogate measures of insulin resistance compared to the hyperinsulinemic euglycemic glucose clamp, log(HOMA-IR) values were shown to be well correlated with measurements of insulin resistance obtained using the glucose clamp (r=−0.60 [−0.66, −0.53], n=22). The conclusion was that in studies using single, paired measurements of fasting blood

glucose and insulin as input data, log(HOMA-IR) would be an appropriate choice for estimating insulin resistance (185).

2.8 Blood Sample Collection, Processing and Storage

All researchers handling samples collected for The Dad's Health Study were required to have completed the Good Clinical Practice eLearning (Secondary Care) course with a satisfactory score (186). Further, all individuals collecting samples for The Dad's Health Study were required to have completed the Medical Research Council's (MRC) Research and Human Tissue Legislation e-learning Package with a satisfactory score on the online module assessment (187). Tissue samples were handled in accordance with guidelines set out in the MRC 'Use of human samples in medical research' (188). Following these guidelines ensured that samples were handled in accordance with the 2004 Human Tissue Act.

Blood samples (parental and cord blood samples) were ultracentrifuged in the same laboratory within the Fetal Medicine Unit (FMU) in UCLH. The samples were aliquoted as described in section 2.6 and the vials labelled with their unique participant identification number, date of sample provision and type of sample, before being transported over the UCL EGA Institute for Women's Health Chenies Mews laboratory in plastic, airtight containers. DNA extraction was performed in the UCL EGA Institute for Women's Health Chenies Mews laboratory using safety precautions detailed in the MRC 'Use of human samples in medical research' guidelines (187). Following DNA extraction, samples were stored in anonymised form in the designated −80°C freezer in the UCL Paul O'Gorman Building, which has restricted access.

2.8.1 DNA Extraction from Blood Samples

DNA was extracted from buffy coat in the same manner for parental and cord blood samples using 200 μL buffy coat according to instructions in the *'DNA Purification from Blood or Body Fluids (Spin Protocol)'* handbook from the *QIAamp® DNA Mini Kit* (180). All centrifugation steps were carried out in room temperature. Prior to

commencing each DNA extraction, it was ensured that appropriate amounts of 96-100% ethanol had been added to relevant Qiagen buffers.

First, 20 μL of Qiagen Protease was pipetted into the bottom of a 1.5 mL microcentrifuge tube. Qiagen Protease is a serine protease that functions to digest a broad range of potentially contaminating proteins. It has an increased activity at higher temperatures. 200 μL buffy coat was then added to the microcentrifuge tube. As RNA-free DNA was not required for later experiments, the optional step of adding RNase A stock solution was omitted. 200 μL of Qiagen Buffer AL (lysis buffer) was then added to the microcentrifuge tube. While the exact composition of Qiagen Buffer AL remains proprietary, the lysis buffer contains guanidine hydrochloride, a chaotropic salt denaturant that functions to dissociate nucleoproteins and inhibit nucleases (189). The microcentrifuge tubes containing Qiagen Protease, the buffy coat sample and buffer AL was then pulse-vortexed for 15 seconds to ensure mixing, and then incubated in a 56° C water bath for 10 minutes. This temperature both works to increase the activity of Qiagen Protease as well as helps to denature DNases and RNases, ensuring that nucleic acids present in the tissue sample are left intact for later purification.

Following incubation, the microcentrifuge tube was centrifuged at 13,000g for 15 seconds to remove drops from inside the lid. 200 μL ethanol (>96%) was added, and the microcentrifuge contents were mixed by pulse vortexing for 15 seconds. In the presence of a buffer with a high concentration of chaotropic salt, such as Buffer AL, and a low (<7) pH, ethanol causes precipitation of DNA. The mixture was carefully pipetted onto a QIAamp Mini spin column in a 2 mL collection tube, making sure not the wet the rim, and centrifuged at 13,000g for one minute. During centrifugation, the DNA adsorbs specifically to the QIAamp silica-gel membrane while contaminants pass through. The collection tube containing the filtrate was discarded and the QIAamp Mini spin column was placed in a fresh 2 mL collection tube.

The QIAamp Mini spin column was then opened and 500 μL of Qiagen Buffer AW1 (wash buffer 1) was pipetted onto the column membrane. The cap was closed and

the column (in the 2 mL collection tube) was ultracentrifuged at 6,000g for 1 minute. While the exact composition of Qiagen Buffer AW1 remains proprietary, it contains a low concentration of chaotropic salts (e.g. guanidine hydrochloride) that help to remove residual protein from the QIAamp Mini spin column membrane. Following centrifugation, the collection tube containing the filtrate was discarded and the QIAamp Mini spin column was placed in a fresh 2 mL collection tube. The QIAamp Mini spin column was then opened and 500 μL of Qiagen Buffer AW2 (wash buffer 2) was pipetted onto the column membrane. The cap was closed and the column (in the 2 mL collection tube) was ultracentrifuged at 13,000 g for 3 minutes. While the exact composition of Qiagen Buffer AW2 remains proprietary, it contains a tris-ethanol solution that helps to remove residual salts from the QIAamp Mini spin column membrane. Following centrifugation, the collection tube containing the filtrate was discarded and the QIAamp Mini spin column was placed in a fresh 1.5 mL microcentrifuge tube.

The QIAamp Mini spin column was then opened and 200 μL of Qiagen Buffer AE (elution buffer) was added onto the spin column membrane. The mixture was incubated at room temperature for 5 minutes. During this elution step, the DNA on the spin column silica-gel membrane dissolves in the elution buffer, which contains 10 mM Tris-Cl and 0.5 mM EDTA, and has a pH of 9.0 (190). The slightly basic pH of the buffer allows the DNA to dissolve more rapidly. Following incubation, the QIAamp Mini spin column, in the 1.5 mL microcentrifuge tube, were ultracentrifuged at 4,000g for 1 minute. The spin column was discarded and the microcentrifuge tube containing the eluted DNA was labelled, quality controlled as described in section 2.8.2 and stored as described in section 2.8.

2.8.2 Quality Control of DNA Extracted from Blood Samples

The quantity of DNA obtained from blood samples was measured using a Qubit 3.0 Fluorometer (Life Technologies) and the Qubit dsDNA broad range assay kit according to manufacturer's instructions. DNA yields were considered adequate if they were in the range of 25-50 μg, corresponding to the expected DNA yield detailed in the *'DNA*

Purification from Blood or Body Fluids (Spin Protocol)' handbook from the *QIAamp® DNA Mini Kit* (180).

2.9 Studies of DNA Methylation in Human Sperm: A Systematic Review

A systematic review of studies investigating DNA methylation in human sperm was undertaken to summarise current knowledge, generate recommendations for research and inform the next stages of my project. The methods employed for the systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (191).

2.9.1 Systematic Review Search Strategy

An exhaustive literature search was performed using the PubMed electronic database on March 31st 2019. The following search criteria was applied:

"semen" [Mesh] OR "sperm" [Mesh] AND "DNA methylation" [Mesh] OR "methylome" [Mesh]

The same search criteria were thereafter applied to search the Web of Science and Cochrane databases in order to identify further relevant articles. Only publications that pertained to studies in humans and were written in English were included. Reference lists of identified articles and review articles were also searched for additional references.

The primary search and filtering of search results was performed by myself. The search process was thereafter repeated by an independent researcher. Any differences in the selection of articles between myself and the independent researcher were resolved by discussion.

2.9.2 Selection Criteria for Inclusion of Articles

Articles were selected for inclusion in the systematic review if they fulfilled the following criteria:

- Studies investigating DNA methylation in human spermatozoa
- Any age group
- Investigations of any type of analysis of DNA methylation
- Publications in English only
- Published from 2003 onwards

Initially, publications relevant to the topic of interest were included irrespective of the year of their publication. However, upon later discussion regarding significant progress in DNA methylation analysis technologies that have occurred in particular in the last two decades, a cut-off of 2003 was decided. With the completion of the sequencing of the human genome in 2003, the use of Next Generation Sequencing technologies became more widespread, even though the first human second generation (short read) human genome was not introduced until 2008 (192, 193). Studies performed prior to 2003 typically analysed DNA methylation by restriction endonuclease digestion followed by Southern blotting. We deemed that these earlier studies, performed prior to 2003, employed methods of assaying DNA methylation that were too different from more recent methods to not make results comparable across studies.

Articles identified through the search, but which did not meet the inclusion criteria detailed above were recorded, along with their reason for being excluded.

2.9.3 Data Extraction and Evaluation of Quality of Evidence

The following information was extracted from every identified study:

- Author
- Year of Publication
- Rationale for study
- Method of assaying DNA methylation
- Area of research:
    - Fertility/ Assisted Reproductive Technologies
    - Aging

    - Toxins
    - Diet/ lifestyle/ metabolic disease
    - Cancer
    - Neurological disease
    - Methodology
    - Tissue specificity
    - Other
- Method of semen processing
- Targeted, global or genome-wide analysis
- Number of study participants
- Age of study participants
- Matching of cases and controls (where relevant)
- Main results of study
- Statistical analyses, e.g. correction for multiple comparison

The GRADE criteria were used to objectively evaluate the quality of evidence in every study included in the systematic review (194). The criteria were adapted for relevance to the subject of the review, as summarised in Table 2-1 and discussed in detail below.

GRADE criteria	Rating	Reasons for down- or upgrading	Quality of the evidence
Study design	RCT (High) Non-RCT (Low)	Only non-RCTs	High Moderate Low Very low
Risk of Bias	No Serious (-1) Very serious (-2)	Age Smoking Somatic cell contamination Storage time/sample storage conditions Medication/supplement use Drug and alcohol use Abstinence Medical history of participants BMI/ diet	
Inconsistency	No Serious (-1) Very serious (-2)	Lack of replication cohort Similar studies showing contradicting results	
Indirectness	No Serious (-1) Very serious (-2)	Generalisability Cross-sectional/longitudinal	
Imprecision	No Serious (-1) Very serious (-2)	Small sample sizes Lack of appropriate controls Varying sperm processing Definition of subfertility Lack of descriptive statistics of participants	
Publication Bias	Undetected Strongly suspected (-1)	Candidate gene approaches	
Other (upgrading factors)	Large effect (+1 or +2) Dose response (+1 or +2) No plausible confounding (+1 or +2)		

Table 2-1. GRADE criteria as adapted for the systematic review of studies of DNA methylation in human sperm.
RCT = Randomised Controlled Trial. Adapted from (194).

Study design: The highest scoring studies would have been randomised controlled trials (RCTs). However, no such studies were identified in the search and therefore studies were mainly graded based on other factors.

Risk of Bias: The grade for risk of bias was primarily based on whether appropriate confounders had been accounted for. In particular, I assessed whether age and smoking status of participants had been taken into account, as these factors are known to significantly impact on DNA methylation (195, 196). Also, the potential degree of somatic cell contamination, i.e. whether samples had been adequately processed to remove cells such as leukocytes and epithelial cells, strongly influenced the scoring in terms of risk of bias. It is generally agreed that studies should be performed on the healthy, motile portions of the spermatozoa, which would be more representative of those leading to fertilisation and therefore of more relevance to offspring development.

There are several methods of selecting for motile spermatozoa and cleaning the semen sample, referred to as 'initial processing' in the systematic review. The most common methods include a swim-up procedure and a discontinuous gradient ultracentrifugation procedure. In the former, the semen sample is layered under a sperm wash medium, allowed to incubate, and then the motile spermatozoa are harvested from the supernatant. In the latter, the semen sample is layered on top of two (or sometimes three) layers of different media, ultracentrifuged, and all but the pellet of motile spermatozoa is removed (Figure 2-4). Media used for the gradient centrifugation procedure include e.g. PureSperm, Percoll or Isolate media. Following selection of motile spermatozoa, the sample should be assessed for any remaining somatic cells, e.g. microscopically. Somatic cell contamination can also be adjusted for bioinformatically.

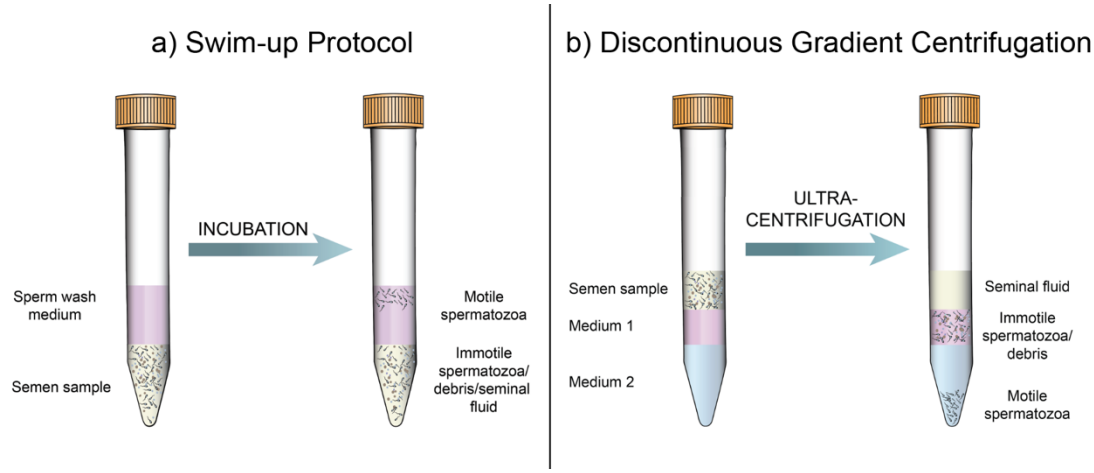


Figure 2-3. Common methods of selecting motile spermatozoa

A score of -2 was given if a study had taken no or little account of confounders and somatic cell contamination, and a score of -1 was given if a study had taken some, but insufficient, account of confounders and somatic cell contamination.

Inconsistency: Inconsistency was graded based on whether similar studies had shown contradicting results and whether the study had included a replication cohort. A score of -2 was given if a study had not included a replication cohort nor replicated results of a previous study, and a score of -1 was given if a study had either replicated previous findings or included a replication cohort.

Indirectness: Indirectness was graded based on the generalisability of the findings, and whether studies were cross-sectional versus longitudinal. Factors that were particularly assessed was whether participants had been recruited from the general public or from reproductive medicine settings without taking this into account when discussing findings of the study. A score of -2 was given if a study was cross-sectional and the results had low generalisability, and a score of -1 was given if the study was either longitudinal or was more likely to be generalisable.

Imprecision: Imprecision was graded based on sample size (or if a power calculation had been used to determine an appropriate sample size), presence or absence of an appropriate control population (where relevant) and whether descriptive statistics of

participants was provided. The method of initial sperm processing (for removal of somatic cells) was also taken into account- if no selection method had been employed, the risk of somatic cell contamination was deemed to be high and thus the findings less precisely relevant to spermatozoa. A score of -2 was given if the study satisfied none or few of these criteria, and a score of -1 was given if the study satisfied some of these criteria.

Publication bias: Publication bias was particularly relevant for candidate gene analyses. As these studies are relatively cost-effective and often small-scale, it is possible that studies yielding negative results would have been less likely to be published than studies yielding positive findings. Therefore, candidate gene approaches were given -1 as a baseline for publication bias.

Upgrading Factors: Upgrading factors, i.e. factors that would provide a higher score to studies, included if a dose-response effect was identified, that there was particularly large effect detected or if several confounders had been appropriately taken into account.

After evaluating each study based on the above criteria, studies were given an overall rating: high/moderate/low/very low.

2.9.4 Compiling Results from the Systematic Review

Identified studies were divided into three categories based on their methodology (analyses of global DNA methylation, candidate gene approaches and genome-wide approaches). Conclusions from studies in the three categories were drawn based on the quality of evidence for the specific studies and used to summarise current understanding of DNA methylation in human spermatozoa. Conclusions from the grading of studies discussed in 2.9.3 were used to generate recommendations for future research. Also, conclusions from the systematic review were used to inform the next stages of my project, i.e. to perform an unbiased, genome-wide profiling of matched samples of sperm and blood from healthy, fertile males.

2.10 DNA methylation Covariation in Spermatozoa and Blood

In order to investigate whether any observed association between paternal metabolic health and offspring birth weight could be mediated by spermatozoal DNA methylation changes passed from the father to his offspring, paternal semen was collected from consenting fathers taking part in the Dad's Health Study. The initial analysis, processing, DNA extraction and DNA methylation analysis of these semen samples is described in sections 2.10 to 2.12.6.

2.10.1 Semen Sample Collection

Participants who provided informed consent to the provision of semen samples as part of The Dad's Health Study were given the choice of producing the sample at home and bringing it to the UCLH Fertility and Reproductive Medicine Laboratory within one hour of sample production, or to produce the sample in the designated rooms in the Reproductive Medicine Unit. Most participants provided the semen sample within a month following their study visit, i.e. within the first half of their partner's pregnancy. All semen samples were processed within one hour of sample production in line with standard guidelines followed by the UCLH Fertility and Reproductive Medicine Laboratory. Participants were not required to abstain prior to sample provision.

As per ethical guidance obtained from the South East Coast - Surrey Research Ethics Committee prior to commencing study recruitment, no semen test results were discussed with participants due to the potential issue of non-paternity.

2.10.2 Semen Sample Handling and Storage

Semen samples collected as part of The Dad's Health Study were handled in accordance with guidelines set out in the MRC 'Use of human samples in medical research' (187). Following these guidelines ensured that samples were handled in accordance with the 2004 Human Tissue Act.

Researchers handling semen samples collected as part of The Dad's Health Study (myself and Research Midwife Anna Greco) underwent training by senior clinical

andrologist Elizabeth Williamson to ensure that samples were handled and processed according to standard operating procedures employed in the UCLH Fertility and Reproductive Medicine Laboratory.

Following the initial analysis and processing of semen samples described in sections 2.10.3 and 2.10.4, samples were labelled with their unique participant identification number and date of sample provision before being transported over the UCL EGA Institute for Women's Health Chenies Mews laboratory in plastic, airtight containers. DNA extraction, as detailed in section 2.10.5, was performed in the UCL EGA Institute for Women's Health Chenies Mews laboratory using safety precautions detailed in the MRC 'Use of human samples in medical research' guidelines (187). Following DNA extraction, samples were stored in anonymised form in the designated –80°C freezer in the UCL Paul O'Gorman Building, which has restricted access.

2.10.3 Initial Analysis of Semen Samples

Within one hour of sample production, samples were placed on a warm plate (35-37°C) for 20-30 minutes in order to liquefy prior to processing. Sample volume was measured and 5µL of the sample was pipetted onto a Leja disposable counting chamber and analysed using the Computer-Assisted Sperm Analysis (CASA)/Sperminator software (Pro-Creative Diagnostics, Staffordshire, UK). The semen sample parameters measured were sperm concentration (millions/mL), percentages of sperm in four categories of motility from most motile to least motile (A – D), and average motile speed (microns/second).

2.10.4 Initial Processing of Semen Samples

Following the initial analysis of semen samples as described in section 2.10.3, samples were processed to select for the more motile sperm and clean them from seminal fluid, cellular debris, epithelial cells and leukocytes. The process is illustrated in Figure 2-4 and described in detail below.

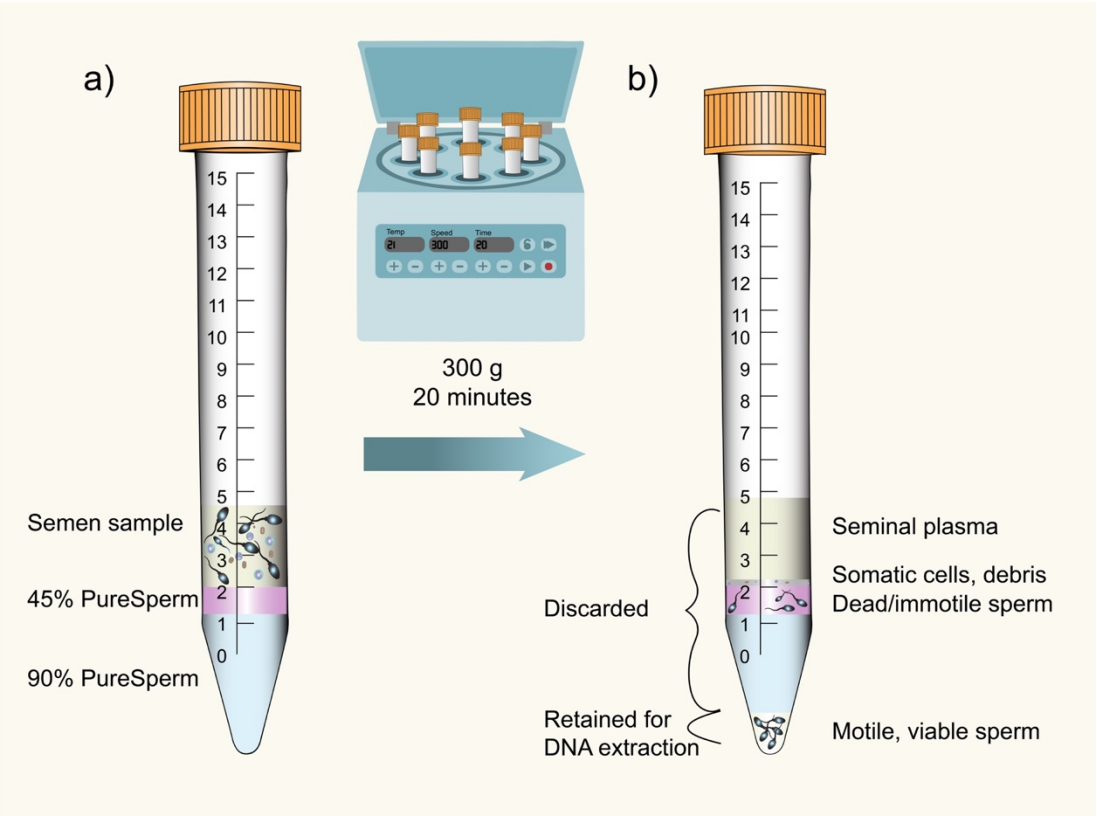


Figure 2-4. Initial processing of semen samples.
a) Semen samples were gently layered on top of a gradient consisting of 1 ml 45% PureSperm above 1 ml 90% PureSperm in order to select for motile sperm (see text for details). b) Following ultracentrifugation, motile, viable sperm form a pellet at the bottom of the conical tube.

Thus, a density gradient was set up in a Corning® 15mL centrifuge tube; 1 ml liquid consisting of 90% PureSperm (PureSperm 100®, Nidacon Laboratories) solution and 10% sperm washing medium (Quinn's™ Sperm Washing Medium, Origio Laboratories) was gently layered under 1 ml liquid consisting of 45% PureSperm and 55% sperm washing medium, while ensuring that the interface between the layers was undisturbed. The liquefied semen sample was gently layered on top of the 45% PureSperm solution and then centrifuged at 300g for 20 minutes. Following this, the supernatant was removed, leaving only the pellet of motile sperm in the bottom of the conical tube. The pellet was then re-suspended in 1 ml sperm washing medium in a clean 15mL centrifuge tube and centrifuged at 300g for 10 minutes. Following this, the supernatant was again removed, and the pellet was re-suspended in 1 ml sperm washing medium in a clean 15mL centrifuge tube. 5µL of the washed sample was loaded onto a Leja disposable counting chamber and analysed using the

Computer-Assisted sperm analysis (CASA)/Sperminator software, assessing the sperm concentration, motility and average motile speed as described above.

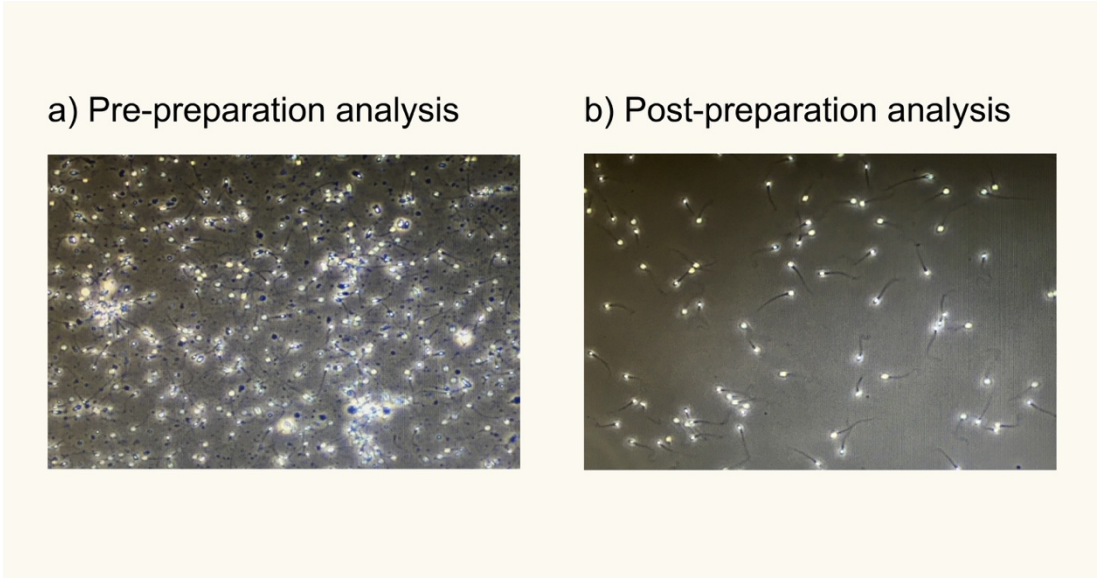


Figure 2-5. Example of semen prior to and after selection of motile sperm.
a) Prior to selection of motile spermatozoa
b) Following selection of motile spermatozoa
Images are taken at the same resolution using the Computer-Assisted sperm analysis (CASA)/Sperminator software. In a) there is visible debris and contaminating cells whereas in b) the samples has been cleaned of contaminating cells as well as dead/immotile semen

2.10.5 DNA Extraction from Semen Samples

The protocol for DNA extraction from semen was obtained from Dr Michelle Holland, Queen Mary University London (QMUL), and had previously successfully been applied to DNA extraction from mouse semen. Its applicability to human semen samples collected as part of The Dad's Health Study had was evaluated in the Pilot Study discussed in section 2.3.2.

Samples were first pelleted in a microcentrifuge tube (Eppendorf, Germany) by ultracentrifugation for 10 minutes at 300g. 250 µL proteinase K buffer (10 mM Tris-HCL, 100 mM NaCl, 25 mM EDTA and 1% SDS), 2.5 µL 1M dithiothreitol (DTT) and 5 µL proteinase K was added to the pellet, and samples were incubated for 5-12 hours in a 55°C water bath (until completely lysed). 250 µL phenol was added to the samples and then shaken vigorously by hand for 5 minutes before centrifuging at 13,000g for

5 minutes. 225 µL of the top aqueous phase was placed in a fresh 2.0 mL microcentrifuge tube to which 225 µL phenol-chloroform was added. This was again shaken for 5 minutes and then centrifuged at 13,000g for 5 minutes. 210 µL of the top aqueous phase was placed in a fresh 2.0 mL microcentrifuge tube to which 210 µL chloroform was added. This was again shaken for 5 minutes and then centrifuged at 13,000g for 5 minutes. 200 µL of the top aqueous phase was placed in a fresh 2.0 mL microcentrifuge tube. 20 µL 3M sodium acetate and 500 µL 96-100% ethanol was added. Upon gentle inversion, the DNA precipitated. Samples were centrifuged for 15 minutes at 13,000g; the supernatant was removed, and DNA pellets were re-suspended in 500 µL 70% ethanol. Samples were again centrifuged for 15 minutes at 13,000g; the supernatant was removed, and DNA pellets were finally suspended in 50 µL Tris-EDTA buffer before being placed in -80°C until further processing and analysis.

2.10.6 Quality Control of DNA Extracted from Semen Samples

The quantity of DNA obtained from semen samples was measured using a Qubit 3.0 Fluorometer (Life Technologies) and the Qubit dsDNA broad range assay kit according to manufacturer's instructions.

2.11 Bisulfite Conversion of Sperm- and Blood- derived DNA samples

Bisulfite conversion of DNA from blood and semen samples was performed using the D5001 EZ DNA Methylation Kit (Zymo Research, Irvine, CA, USA) according to manufacturer's instructions. This kit employs a three-step reaction procedure that is based on the fact that sodium bisulfite converts unmethylated cytosines (C) to uracil (U), while leaving methylated cytosines unchanged. 500 ng (+/-10%) of sample DNA quantified with a Qubit 3.0 Fluorometer (Life Technologies) was used as input DNA.

First, the M-Wash buffer was prepared. Thus, 24 mL ethanol, (≥99.8% ethanol, VWR Chemicals, United Kingdom) was added to the M-Wash buffer concentrate (supplied in the kit) and the and the bottle was inverted several times to allow mixing.

Next, the CT Conversion agent, the powder form of which is supplied in the kit, was prepared. Thus, 750 µL water for molecular biology (Millipore, Germany) and 210 µL of M-dilution buffer (supplied in the kit) were added to the CT conversion reagent. The mixture was vortexed frequently during a period of 10 minutes in order to allow the CT conversion reagent to dissolve. The prepared CT conversion reagent was used immediately following preparation due to it being light sensitive and could thus otherwise have degraded. The CT Conversion reagent supplied in the kit contains sodium metabisulfite ($Na_2S_2O_5$), which upon addition of water is converted to sodium bisulfite ($NaHSO_3$).

5 µL of M-Dilution buffer was then added to the DNA samples, and the total volume was adjusted to 50 µL with water for molecular biology (Millipore, Germany). The samples were then incubated for 15 minutes at 37 °C on a hot plate. The M-Dilution buffer contains sodium hydroxide and aims to denature the DNA in order to make it accessible to the sodium bisulfite for conversion of unmethylated cytosines to uracil. Following the incubation, 100 µL of the prepared CT Conversion Reagent was added to each sample, and the mixture was briefly vortexed. The samples were then incubated in a thermocycler under the following conditions:
(95 °C for 30 seconds, 50 °C for 60 minutes) x 16 cycles, after which they were held at 4 °C.

Following the incubation above, 400 µL of M-Binding buffer (supplied in the kit) was added to a Zymo-Spin IC Column (supplied in the kit) and the column was placed into a collection tube (supplied in the kit). The samples were loaded into the Zymo-Spin IC Columns containing the M-Binding buffer, the cap was closed and the column (with the collection tube) was inverted several times to allow mixing. The columns (with the collection tubes) were ultracentrifuged at 13,000g for 30 seconds after which the flow-through was discarded. 100 µL of the prepared M-Wash buffer was added to the column, and the columns (with the collection tubes) were ultracentrifuged at 13,000g for 30 seconds. After this, 200 µL of M-Desulphonation buffer (supplied in the kit) was added to the column and the columns were left to incubate at room temperature (25°C) for 20 minutes. Following the incubation, the columns (with the

collection tubes) were ultracentrifuged at 13,000g for 30 seconds. The flow-through was discarded, and 200 µL of the prepared M-Wash buffer was added to the column. The columns (with the collection tubes) were then ultracentrifuged at 13,000g for 30 seconds. This wash step was then repeated such that 200 µL of the prepared M-Wash buffer was again added to the column and then the columns (with the collection tubes) were ultracentrifuged at 13,000g for 30 seconds. The flow-through was discarded.

Next, the columns were placed into 1.5 mL microcentrifuge tubes (Eppendorf, Germany). 10 µL of M-Elution buffer was added onto the column matrix and the columns (with the collection tubes) were ultracentrifuged at 13,000g for 30 seconds to elute the DNA.

The bisulfite converted DNA (bsDNA) was stored at -20 °C for a limited time (2-3 days) before analysis using the Illumina MethylationEPIC Array as described below.

2.12 DNA Methylation Analysis in Blood and Semen

For an overview of techniques for DNA methylation profiling, please see section 1.7.1.

Genome-wide CpG methylation for the study of DNA methylation in sperm and blood was measured using the Illumina MethylationEPIC Array. Introduced in 2015, this is the most recent array-based DNA methylation assay, and was specifically designed to interrogate potential enhancer regions as well as 90% or the CpG sites assayed by its precursor, the HM450 (197). The EPIC array assays over 850,000 CpG sites, 350,000 of which are in enhancer regions of the genome. Overall, the EPIC array interrogates >95% of CpG sites in CpG islands, >90% of CpG sites in North and South shores and >80% of CpG sites in North and South shelves as well as gene bodies, potential enhancers and promoter regions (see Figure 2-6).

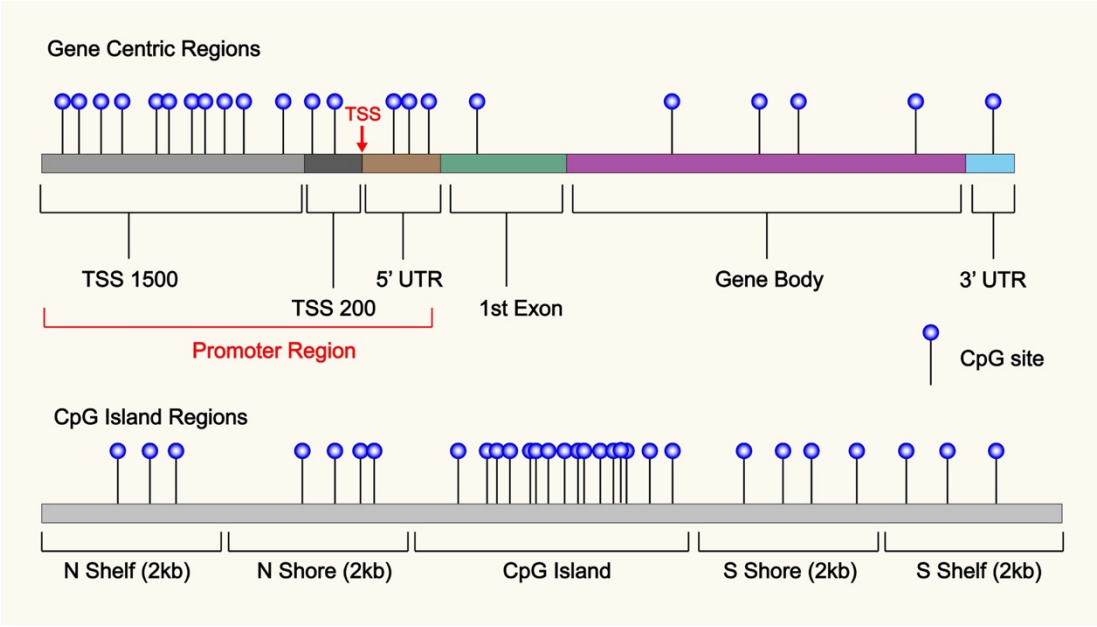


Figure 2-6. Relative density distribution of CpG sites in genomic regions assayed by the EPIC array. The EPIC array interrogates >95% of CpG sites in CpG islands, >90% of CpG sites in North and South shores and >80% of CpG sites in North and South shelves, in addition to high coverage of potential enhancer regions, promoter regions and gene bodies. TSS 1500 and TSS 200 denote regions 1500 and 200 base pairs (bp) upstream of the transcription start site respectively. TSS = Transcription Start Site, UTR = Untranslated Region, N = North, S = South. Modified from (198) with information from (197, 199).

2.12.1 Illumina MethylationEPIC Array Workflow

The overall Illumina EPIC workflow is summarised in Figure 2-7. First, genomic DNA is treated with sodium bisulfite, which converts unmethylated cytosine (C) to uracil (U) while leaving methylated cytosines unchanged (197).

Next, the bisulfite converted DNA is subjected to whole-genome amplification (WGA), during which U is amplified as thymine (T). The DNA is then enzymatically fragmented and prepared for hybridisation to Illumina BeadChips. 12 samples are loaded onto each BeadChip, which have seals to separate the respective samples. Up to 8 BeadChips are incorporated into one kit, allowing analysis of up to 96 samples. The bisulfite converted, amplified and fragmented DNA samples are incubated overnight to allow hybridisation to the BeadChips (200).

The BeadChips contain over 850,000 bead types, each of which contain a locus-specific 50mer probe sequence with a CpG site at the 3' end. The 50mer sequences are designed to be complementary to specific 50 base pair regions of the bisulfite

converted, fragmented DNA. Following hybridisation, single base extension incorporates either a fluorescently labelled biotin nucleotide or a fluorescently labelled dinitrophenyl (DNP) nucleotide; C and G nucleotides are biotin labelled whereas A and T are DNP labelled. The fluorescent signal is then measured (87). To analyse the methylation status of individual CpG sites across the genome, two different types of assay are employed; the Infinium Type I and Infinium Type II assays (see Figure 2-8).

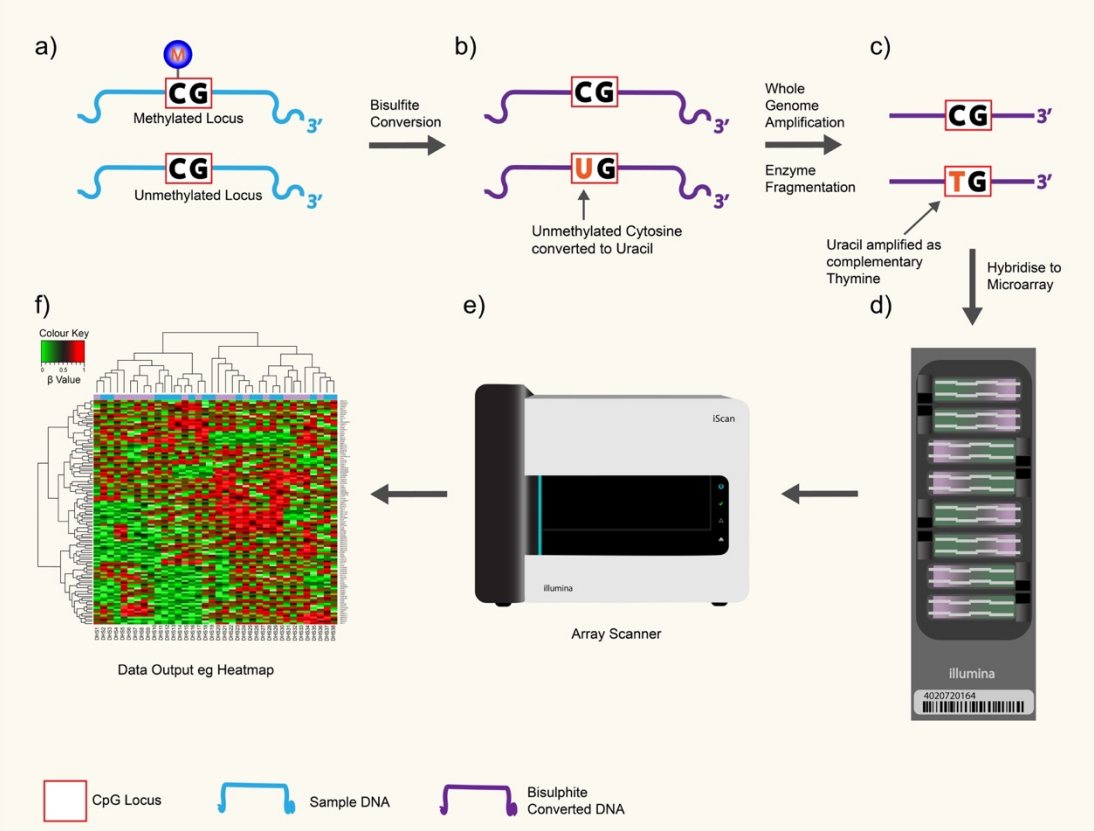


Figure 2-7. Overview of the Illumina MethylationEPIC Array Workflow.
a) Sample DNA with one locus containing a methylated C and one locus containing an unmethylated C. b) The sample DNA is treated with sodium bisulfite, converting unmethylated C to U while leaving methylated C intact. c) The sample DNA undergoes whole genome amplification, during which U is amplified as T. d) The bisulfite-converted, amplified and fragmented DNA is hybridised to the Illumina BeadChips. e) The BeadChips are scanned by the Illumina 'iScan' array scanner. e) Methylation data output in the form of a heat map. Figure created using information from (200) C = cytosine, U = uracil, T = thymine, G = guanine.

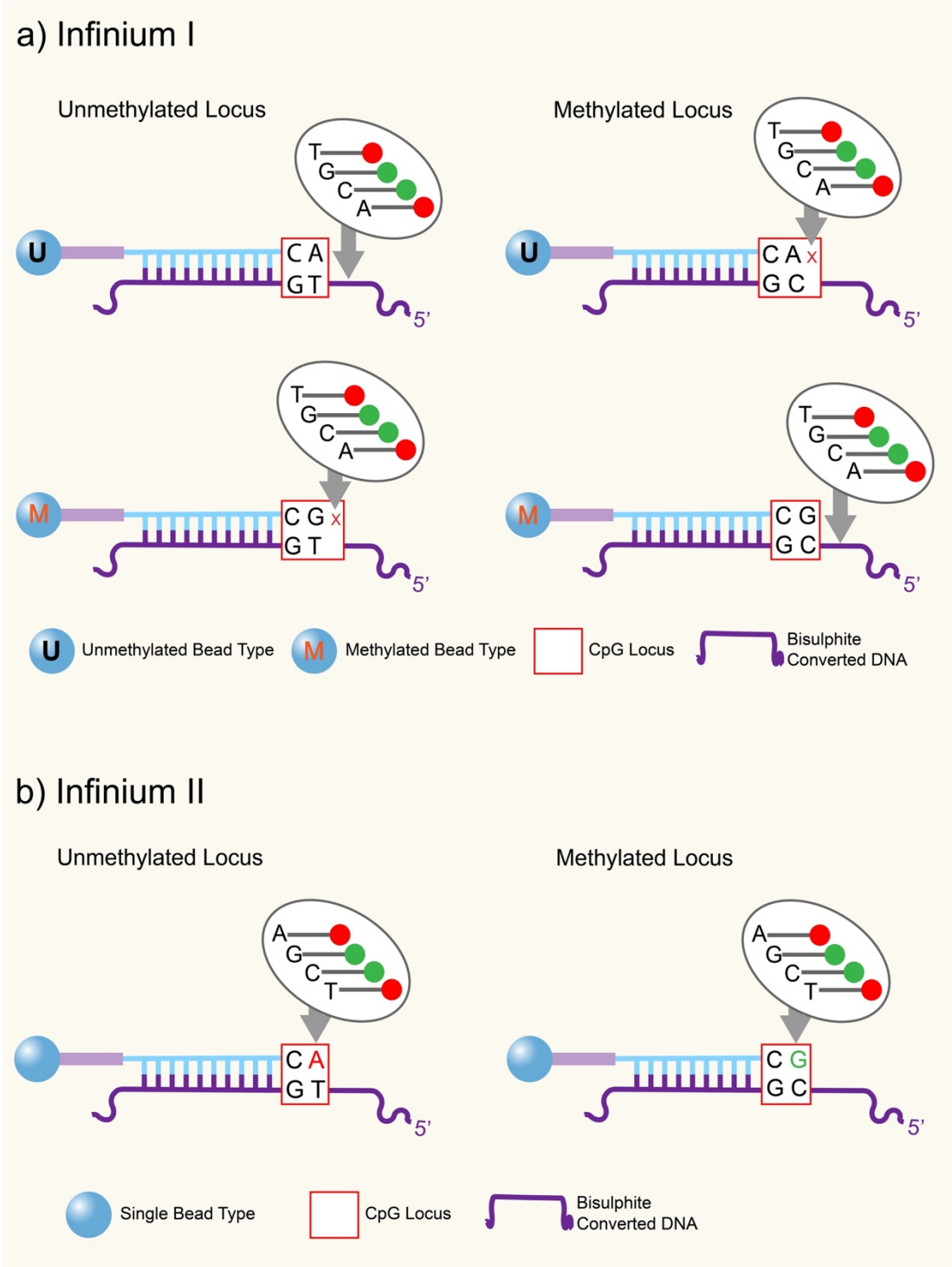


Figure 2-8. The Infinium Type I and Type II assays used to determine CpG methylation status in the MethylationEPIC Array.
a) The Infinium Type I Assay employs two bead types for each CpG sites; one for the methylated (C) and one for the unmethylated (T) state of the site. b) The Infinium Type I Assay employs only one bead type per CpG, and the methylation status is instead determined by the single base extension step following hybridisation. Reproduced from (201).

The Type I assay employs two bead types for each CpG site; the one for the methylated (C), and another for the unmethylated (T) state of the site. This assay relies on the assumption that all CpG sites within a 50 base-pair span share the same methylation status, which is supported be previous research (201). The 50mer probe sequence is thus designed to match either the methylated or the unmethylated version of the CpG site in the sample DNA, and hybridises accordingly. A sample DNA fragment with an unmethylated target locus will thus hybridise to the unmethylated bead type and allow subsequent single base extension of a fluorescently labelled nucleotide matching the nucleotide immediately upstream of the target CpG site. When a sample DNA fragment with an unmethylated target locus hybridises to a methylated probe sequence, however, base extension is inhibited. Vice versa occurs for methylated loci (87).

The Type II assay employs only one type of bead per CpG, and the methylation status of the target CpG site is instead determined by single base extension following hybridisation (see Figure 2-8). For an unmethylated locus, the base complementary to the CpG site in the sample DNA will be adenine (A, complementary to T), whereas for a methylated locus the complementary base will be G (complementary to C) (201). The complementary bases are again fluorescently labelled. In contrast to the Type I assay, the Type II assay design makes no assumptions about the methylation status of neighbouring CpG sites; the other CpG sites within the 50mer sequence are replaced with degenerate R bases that hybridise both to T and C. Further, as one rather than two bead types are employed for each CpG locus, the Type II assay also take up less physical space on the BeadChips. For these reasons, the methylation status of the vast majority of CpG sites covered by the EPIC array (84%) are queried using the Type II probes (197).

Following hybridisation and base extension that incorporates a labelled probe, the hybridised sample DNA is removed. The BeadChips are then imaged using the Illumina iScan System; a laser excites the fluorophore of the single base extension product on the beads and records high resolution images of the BeadChip to visualise the intensities of the methylated and unmethylated bead types (200). The BeadChips

are simultaneously scanned at two wavelengths, the red and the green channel, and an image file is created for each of these. A and T are assessed in the red channel whereas C and G are assessed in the green channel.

2.12.2 Illumina Internal Quality Controls

The Illumina EPIC array includes both sample-dependent and sample-independent controls in order to evaluate the quality of specific steps within the EPIC array process flow as well as performance across samples. There are controls that are assessed in each of the red and green colour channels, and those that are assessed in both (200).

Sample-independent controls include staining controls, extension controls, hybridisation controls and target removal controls (see Figure 2-9). Staining controls assess the efficiency of the single base extension step where a fluorescently labelled nucleotide is incorporated into the probe sequence, and evaluate the efficiency of both the red and the green colour channel. The green channel shows a higher signal for biotin staining when compared to biotin background, whereas the red channel shows higher signal for DNP staining when compared to DNP background. A high level of background fluorescence reduces the ability to distinguish signal from noise and should lead to sample removal.

Extension controls are beads that test the extension efficiency of A, T, C and G nucleotides respectively by using a hairpin probe. Both red (A, T) and green (C, G) are assessed, and a high intensity is expected from the extension control probes if the extension has worked correctly. Hybridisation controls employ synthetic targets that perfectly complement the probe sequences attached to the beads. These synthetic targets come in different levels of concentration (low, medium and high) which should correspond to observed intensities for these probes.

The hybridisation controls should only be assessed in the green channel as biotin rather than DNP is incorporated at the base extension step. Target removal controls assess the efficiency of removing the sample DNA following hybridisation and base extension. The target removal controls are extended using the probe sequence as a

template, and should be efficiently removed to produce a low intensity compared to the hybridisation controls. The target removal controls should only be assessed in the green channel as it is the lack of incorporation of biotin that is assessed.

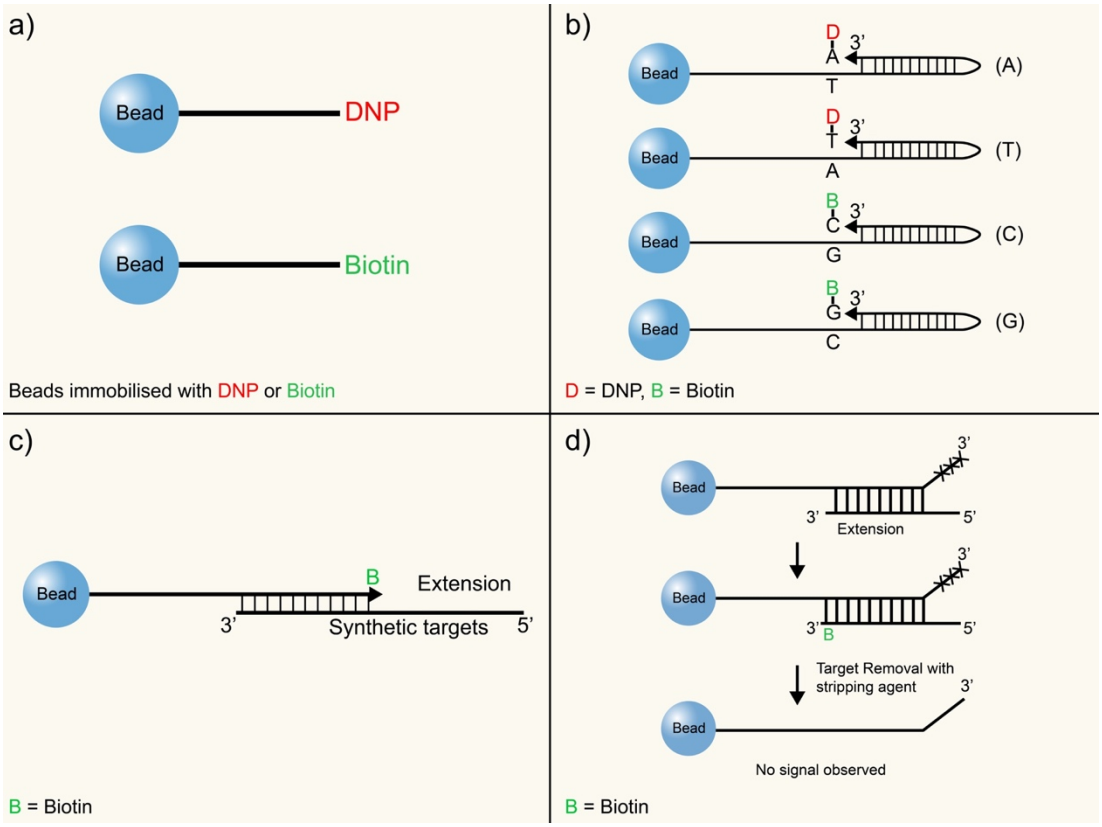


Figure 2-9.Sample-independent controls included in the EPIC array.
a) Staining controls have DNP or Biotin attached to the beads and evaluate the performance of the staining step. b) Extension controls test the efficiency of extension of A, T, C and G from a hairpin probe. c) Hybridisation controls use synthetic controls to test the overall performance of the assay. Synthetic targets are present in low, medium and high concentrations. Target by extension generates a signal, and the signal intensity increases as the target concentration increases. d) Target removal controls test the performance of the stripping step following the extension reaction using a non-extendable probe and a complementary extendable synthetic target. See text for further details. DNP = dinitrophenyl. Modified from (200).

Sample-dependent controls include controls for the efficiency of bisulfite conversion, specificity of probe extension, levels of background intensity and overall performance of the assay (see Figure 2-10).

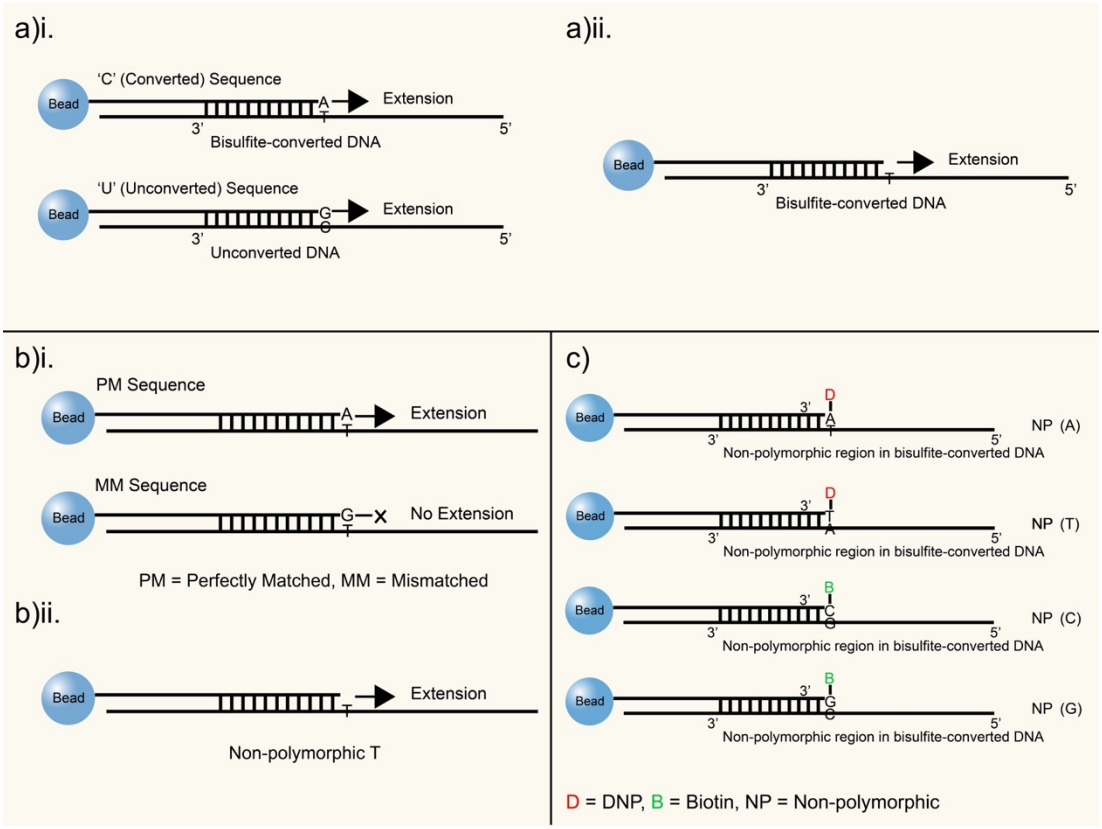


Figure 2-10. Sample-dependent controls for the EPIC array.
a) Bisulfite conversion controls test the efficiency of bisulfite conversion of genomic DNA. a)i. are bisulfite conversion controls that use the Infinium Type I design and a)ii. are bisulfite conversion controls that use the Infinium Type II design b) Specificity controls monitor potential non-specific primer extension for Type I ((b)i.) and Type II ((b)ii.) probes. c) Negative controls are randomly permutated sequences that should not hybridise to the DNA template. See text for further details. DNP = dinitrophenyl. C = Converted, U = Unconverted, PM = Perfectly Matched, MM = Mismatched, D = DNP, B = Biotin, NP = Non-Polymorphic. Modified from (200).

The bisulfite conversion controls assess the efficiency of bisulfite conversion of genomic DNA. There are two types of bisulfite conversion controls; those that employ the Infinium Type I design (Bisulfite Conversion I probes) and those that employ the Infinium Type II design (Bisulfite Conversion II probes). The cytosines queried by the bisulfite conversion control probes are at non-CpG sites and chosen because they are the sole C base in a 50bp span. For Bisulfite Conversion I probes, the 'C' or converted probes will match the converted sequence if the bisulfite conversion was successful, whereas if the sample has unconverted DNA (meaning that the bisulfite conversion was incomplete), the 'U' or unconverted probes will be extended. Bisulfite controls

C1, C2 and C3 should be assessed in the green channel, whereas C4, C5 and C6 should be assessed in the red channel. Bisulfite Conversion II probes will use single base extension to incorporate an A if the bisulfite conversion was successful and a G if the conversion had been unsuccessful. The G would cause an elevated intensity in the green channel.

Specificity controls assess the level of nonspecific probe extension for Infinium Type I (Specificity I) and Infinium Type II (Specificity II) probes, and are designed against nonpolymorphic T sites. If probe extension from Type I probes worked perfectly, then for a methylated CpG locus, the C in the sample DNA (or specificity control) should only anneal to G in the probe sequence. For an unmethylated CpG locus, then the T in the sample DNA (or specificity control) should only anneal to A in the probe sequence. In other words, a G annealing to a T would be a mismatch. Specificity I probes assess the level of G/T mismatch (which prohibits further extension) and compares this to the level of perfect A/T matching (which would allow base extension and yield a high signal compared to a mismatch). Performance of G/T mismatch controls should be assessed in both red and green channels. Specificity II probes should incorporate an A base to complement the nonpolymorphic T and have intensity in the red channel. If there is nonspecific incorporation of G instead, then the probe will have increased intensity in the green channel.

Negative controls define the background intensity for the DNA methylation assay. They are constructed as randomly permutated sequences and thus should not bind specifically to any of the bead types. The mean signal intensity of 600 negative controls is used to establish detection limits for the methylation probes. Performance of negative controls should be monitored in both red and green channels.

Nonpolymorphic controls test the overall performance of the assay, from amplification to detection, by querying a particular base in a nonpolymorphic region of the bisulfite genome. There are four nonpolymorphic controls, one for each of the 4 nucleotides (A, T, C and G). These controls allow comparison of the performance of the assay across different samples.

2.12.3 EPIC Array Data Output

Methylation data at each CpG site is expressed as a beta value (β), which describe the ratio of the methylated (C) to the unmethylated (T) signal, via the formula:

$$\beta = \frac{\text{intensity of methylated signal}}{(\text{intensity of methylated signal} + \text{intensity of unmethylated signal} + 100)}$$

A β value of 0 represents a completely unmethylated CpG site and a β value of 1 represents a fully methylated CpG site. For a single individual, the distribution of β values across all CpG sites is bi-modal, with the majority of CpGs being close to zero or one. However, for a single CpG site across individuals, the distribution of β values tends to be uni-modal (197) . In other words, most of the CpG sites in one individual are either fully methylated or completely unmethylated, and any given CpG site tends to be similarly methylated in one individual compared to another.

The output data from the Illumina iScan System used for analysis of the MethylationEPIC Array is typically presented in the form of an intensity data (IDAT) file, where summary intensities for each probe-type on the array is detailed. IDAT files are binary files, one for each of the red and green channels, and comprises information on four fields; the ID of each bead-type on the array, the mean and standard deviation of their intensities, and the number of beads of each type (202).

2.12.4 Quality Control of DNA Methylation Data from the Illumina MethylationEPIC Array

Before data generated from an EPIC array can be analysed with regards to the outcome variable and its association to CpG methylation, the data needs to undergo a series of quality control and preprocessing steps.

First, samples are checked for inconsistencies in the methylation data. Potential sample replicates that do not correlate are removed and if relevant, samples with a gender mismatch are also removed from further analysis. One way of uncovering sample mismatches is by visualising similarities between samples using a multi-dimensional scaling (MDS) plot. In this, sample clustering can help to identify large

scale similarities between samples and interrogate further should samples deviate from the expected clustering. In studies with samples from both males and females, MDS plots are typically performed to look at methylation on the X chromosome. This should be relatively unmethylated for males and 50% methylated for females (due to X chromosome inactivation). Thus, when using MDS across the X chromosome you would expect samples to cluster according to whether they are from males or females.

Following this, genotyping information can be used to identify potentially mislabelled samples. 59 of the control probes on the EPIC array are for direct interrogation of genotype. If there is separate genotype information for the individuals in a study, these genotype probes can be used cross-checked against existing genotype data. If the samples contain replicates or more than one sample from the same individual, these genotype probes can be used to check whether the genotypes at these 59 probes are identical. If genotypes do not match, this most likely indicates that samples might be mislabelled or otherwise problematic.

2.12.5 Data preprocessing for the Illumina MethylationEPIC Array

Following quality control of the DNA methylation data, outcomes from the various types of internal quality checks and control probes described in section 2.12.2 should be taken into account and also give an indication of the overall reliability of the particular assay. One has to account for the presence of SNPs, cross-reactive probes and the different properties for the Type I and Type II probes described in section 2.12.1. In general, Type II probes have lower reproducibility than Type I probes, and these differences should be normalised prior to analysis of methylation data.

Potentially problematic probes are filtered out. These include the control probes and those probes where a significant proportion (the exact number varies between assays) fail to meet the detection p-value threshold. If a large number of probes for a given sample fail to reach the detection p-value, the entire sample should be removed from further analysis (203).

The presence of SNPs in close proximity to interrogated CpG sites can lead to false assumptions about the association of methylation of a particular CpG site and the outcome variable as SNPs in the probe sequence alter the binding affinity to the probe. For example, if one genotype of the SNP, e.g. A, always comes with a methylated CpG site, and the other genotype, e.g. T, always comes with an unmethylated CpG site and whereas the A genotype might bind the probe sequence 100% of the time, the T may only bind 20% of the time, giving a skewed methylation readout. For this reason, data from e.g. the 1000 Genomes Project or other annotated lists of SNPs can be cross-matched with study results so that these signals can be analysed in more detail and false assumptions can be better avoided (197). Similarly, annotated lists of potentially cross-reactive probes that bind non-specifically to the target regions of interest or that bind to repetitive regions of the genomes, can be used to flag or filter out these probes.

In addition to using annotated lists of probes influenced by SNPs, there are other methods of identifying sites where the CpG methylation level is potentially influenced by the underlying genotype. One of these is “gap hunting”; a method that identifies clustered beta value distributions, such as those consistent with a bi- or trimodal distribution of beta values (204).

The data should then be normalised in order to remove technical and systematic variability to make results comparable across samples. One key aspect of data normalisation for the EPIC array is to make measurements across the two different probe types comparable against one another. There are several ways to achieve this. One approach is to normalise the Type II probes to the Type I probes, so-called BMIQ normalisation, by transforming the distribution of Type II probes to be similar that of Type I probes (197).

2.12.6 Interpreting DNA Methylation data from the MethylationEPIC Array

Once DNA methylation data has been analysed and matched with appropriate study-specific variables (e.g. phenotype profiles or case-control status), inferences can be drawn in terms of which biological processes may be involved. This can be achieved

by performing a gene ontology (GO) analysis, wherein significant CpG sites are annotated to biological networks (205). Pathway analyses can be performed to suggest particular molecular signatures that may be relevant in, for example, a disease process. Such pathways or individual genes can be taken forward for more conclusive investigations, e.g. expression profiling, and in extension may become targets for therapeutic intervention or the development of disease-specific biomarkers.

Results from GO-analyses should, however, be interpreted with caution. For example, some genes represented on the HM450 and the MethylationEPIC bead chips have a considerably higher probe representation than others (206). This introduces a bias as genes with more probes are more likely to appear as differentially methylated. Recent bioinformatic approaches have been developed to overcome this bias. Such an approach includes the empirical Bayes Gene Set Enrichment Analysis (ebGSEA), which directly rank genes according to their overall level of differential methylation as assessed using all of the probes that map to a given gene and in a manner that avoids favouring genes containing more probes, before assessing for enrichment of biological terms using this ranked list of genes (206). Another method of overcoming this problems is to use functions such as the *gometh* function in the missMethyl R package used to analyse methylation data (207). In this, the selection bias is reduced by modelling the relationship between the number of genes per probes and the probability of the gene being selected (e.g. as hyper- or hypomethylated) (207).

Inferences can also be drawn by analysing the genomic locations of identified CpG sites. In very general terms, increased CpG methylation of promoters is associated with transcriptional silencing. CpG methylation changes to enhancer or other regulatory regions can be investigated by cross-referencing these to information compiled by e.g. the ENCODE Project Consortium, which has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification (208). The process of gene regulation of course involves a highly complex interaction between the nucleotide sequence, DNA methylation, small

interfering RNA molecules and histone modifications. Thus, integrating DNA methylation data with information on other genetic and epigenetic processes is vital for providing a more complete understanding of biological processes. For example, continued elucidation of methylation quantitative trait loci (mQTLs) will allow DNA methylation data to be integrated with SNP data to understand how genetic variation may have influenced results derived from genome-wide CpG analyses (209).

2.12.7 DNA methylation analysis using the Gene Expression Omnibus

A further way to interrogate CpG sites is to compare results from a DNA methylation array analysis to large, publicly available methylation data sets. This approach can significantly increase study power and thus increase the probability to detect biologically meaningful results. The MethylationEPIC data from sperm samples in my study was therefore added to data of previously performed DNA methylation (HM45) arrays available on the Gene Expression Omnibus (GEO) database and compared to a wide range of somatic tissues with methylation data available on GEO. The GEO database is a publicly available database that contains Illumina DNA methylation array data from tens of thousands of human tissue samples, as well as other array- and sequencing data (210). Indeed, as on July 2017, data from more than 60,000 HM450 arrays had been deposited on this database (211).

The recently developed bigmelon R package provides a memory efficient workflow to enable these complex, large scale analyses without the requirement for large random access memory (RAM) (211). This package was used to identify CpG sites that displayed hypo- and hypermethylation in sperm when compared to a several types of somatic tissues. Significant CpG sites were therefore investigated in terms of their annotation to biological networks by a GO analysis to identify biological pathways that significantly differed between sperm and somatic tissues.

2.13 Obesity Associated DNA Methylation Profiling in Blood

Validation of previously identified obesity associated CpG sites was performed using a multiplex bisulphite PCR sequencing microfluidics-based assay (Bis-PCR-Seq), which

constitutes a cost-effective, targeted alternative to Reduced Representation Bisulfite Sequencing (RRBS) in genome-wide analyses.

Blood samples collected from 96 lean (BMI 18-25 kg/m^2) and 96 obese (BMI >30kg/m^2) males were included in this validation. Most samples (78 from lean males and 78 from obese males) were collected as part of the Dad's Health study as described in section 2.8. A proportion of blood samples (18 from lean males and 18 from obese males) were collected at the Women's Health Tissue Repository, University of Iowa Health Care, using the same methods. Blood-derived DNA was bisulfite converted as described in section 2.11. CpG sites were identified based on previous EWASs of obesity (section 2.14). Primers were designed to target the specific obesity associated CpG sites and were evaluated for performance (section 2.14.1). Selected primers were then used to validate obesity associated CpG methylation in the blood-derived, bisulfite converted DNA samples. This validation was performed using the Fluidigm Access Array, a multiplex bisulphite PCR sequencing microfluidics-based assay (section 2.15).

This study forms the first stage of a larger project that aims to generate a robust, reproducible obesity associated DNA methylation profile using peripheral blood from 1000 lean and 1000 obese males (MRC reference code MR/P011799/1; title *'Paternal obesity-associated DNA methylation: an investigation into its reproducibility, reversibility and association with fetal growth restriction'*).

2.14 Identification of obesity associated CpG sites

Identification of obesity associated CpG sites for validation was performed through a PubMed search using the following search criteria:

- Analysis of DNA methylation in association to human obesity or BMI
- Minimum sample size of 1000 participants (discovery and replication cohorts combined)
- Performed using the Illumina HM450 array

The search was performed on October 6th, 2017. Therefore, some of the most recent obesity EWASs discussed in section 1.8.1 were not included. In total, this led to the inclusion of six EWASs of obesity/BMI, which are summarised in Table 2-2.

Author, year, reference		*n*	*Phenotype*	*Tissue*	*Number of CpG Sites identified (multiple testing correction)*	*Ethnicity*	*Cohort*
Aslibekyan et al., 2015 (101)	Discovery cohort	991	BMI	CD4+ T cells	8 (Bonf.)	European Americans	GOLDN
	Replication cohort 1	2105	BMI	Whole blood	2 (Bonf.)	European Americans	FHS
	Replication cohort 2	1935	BMI	Whole blood	4 (Bonf.)	African Americans	ARIC
	Meta-analysis	5031	BMI	Whole blood	8		
Demerath et al., 2015 (102)	Discovery cohort	2097	BMI	PBLs	76 (Bonf.)	African Americans	ARIC
	Replication cohort	3368	BMI	Whole blood/ CD4+ T cells	37 (Bonf.)	European Americans	FHS + GOLDN
Al Muftah et al., 2016 (103)	Discovery cohort	123	BMI	Whole blood	None	Arab	Quatari family study
	Replication cohort	810	BMI	Whole blood	None	Caucasian	TwinsUK
	Meta-analysis*	123 + 810	BMI	Whole blood	None		
Mendelson et al., 2017 (104)	Discovery cohort	3743	BMI	Whole blood	135 (Bonf.)	European Americans + Caucasian	FHS + LBCs
	Replication cohort	4055	BMI	Whole blood/CD4+ T cells	83 (Bonf.)	African Americans + European Americans + Caucasian	ARIC + GOLDN + PIVUS
Sayols-Baixeras et al., 2017 (105)	Discovery cohort	641	BMI	Whole blood	94** (Bonf.)	European	REGICOR
	Replication cohort	2515	WC	Whole blood	49** (Bonf.)	European Americans	FOS
Wahl et al., 2017 (106)	Discovery cohort	5387	BMI	Whole blood	278 (Bonf.)	European + Indian-Asian	EPICOR + KORA + LOLIPOP

Replication cohort	4874	BMI	Whole blood	187 (Bonf.)	European + Indian-Asian	ALSPAC + EGCUT + Leiden Longevity + LifeLines Deep + LOLIPOP + RS-BIOS + RS-III + TwinsUK

Table 2-2. EWASs of obesity associated genome-wide CpG methylation included in the validation study.

Note that some of these studies identified CpG sites associated with BMI-related traits, such as waist circumference and T2DM. However, these findings are not presented in the table.

* This study replicated 1 out of 8 T2DM and 7 out of 39 BMI findings from previous EWASs in the Quatari discovery cohort. The meta-analysis was performed on these 8 replicated loci

* Results are from a meta-analysis of discovery and replication cohorts.

WC = Waist Circumference, PBL = Peripheral Blood Leukocyte, FDR = False Discovery Rate, Bonf = Bonferroni, ARIC = Atherosclerosis Risk in Communities, GOLDN = Genetics of Lipid Lowering Drugs and Diet Network, FHS = Framingham Heart Study, REGICOR = Girona Heart Registry, FOS = Framingham Offspring Study, LOLIPOP = London Life Sciences Population Study, ALSPAC = Avon Longitudinal Study of Parents and Children, EGCUT = Estonian Genome Center of the University of Tartu, RS-BIOS = Rotterdam Study Bios Cohort, RS-III = Rotterdam Study-III

One further study was included as it investigated the association between obesity associated SNPs and methylation of nearby CpG sites (212). This study interrogated methylation levels at CpG sites in proximity to 52 obesity associated SNPs, and identified 107 CpG sites in close proximity to 28 SNPs where methylation levels were associated with the genotype. One of these were removed as two CpG sites were associated with the same SNP. In addition, CpG sites where methylation levels have previously been shown to be influenced by epigenetic confounders such as smoking and inflammation (as measured by C-reactive protein, CRP) were included in order to be able to control for these variables. The smoking associated sites were identified from a meta-analysis of EWASs of smoking which included almost 16,000 participants in total (195). The CRP associated sites were identified from a meta-analysis of EWASs of serum C-reactive protein that included almost 13,000 participants in total (213). The blood cell composition control sites were identified from (214). A total of 546 CpG sites were identified using the described criteria.

Prior to primer generation, the CpG sites identified then underwent filtering based on the following criteria:

- Consistency of discovery p value. The empirically derived threshold of $p = 1.16 \times 10^{-7}$ was employed as all studies were performed using the Illumina HM450
- Feasibility of generating a primer for the specific CpG site. This was determined mainly based on the surrounding genetic sequence
- Ability of primers to align uniquely to the human genome
- Ability of primers to capture the specific CpG site, i.e. that the CpG site was situated within 130bp of either primer end. This is to account for the drop off in read quality observed in 150 bp paired-end Illumina MiSeq System reads
- No confounding genetic variation in the primer sequence

Together with the technical requirement of processing samples in batches of 48, the above criteria lead to the removal of 162 CpG sites. Thus, 384 CpG sites were selected for primer generation. The process of identifying and selecting CpG sites for inclusion in the validation of obesity associated CpG sites is summarised in Figure 2-11.

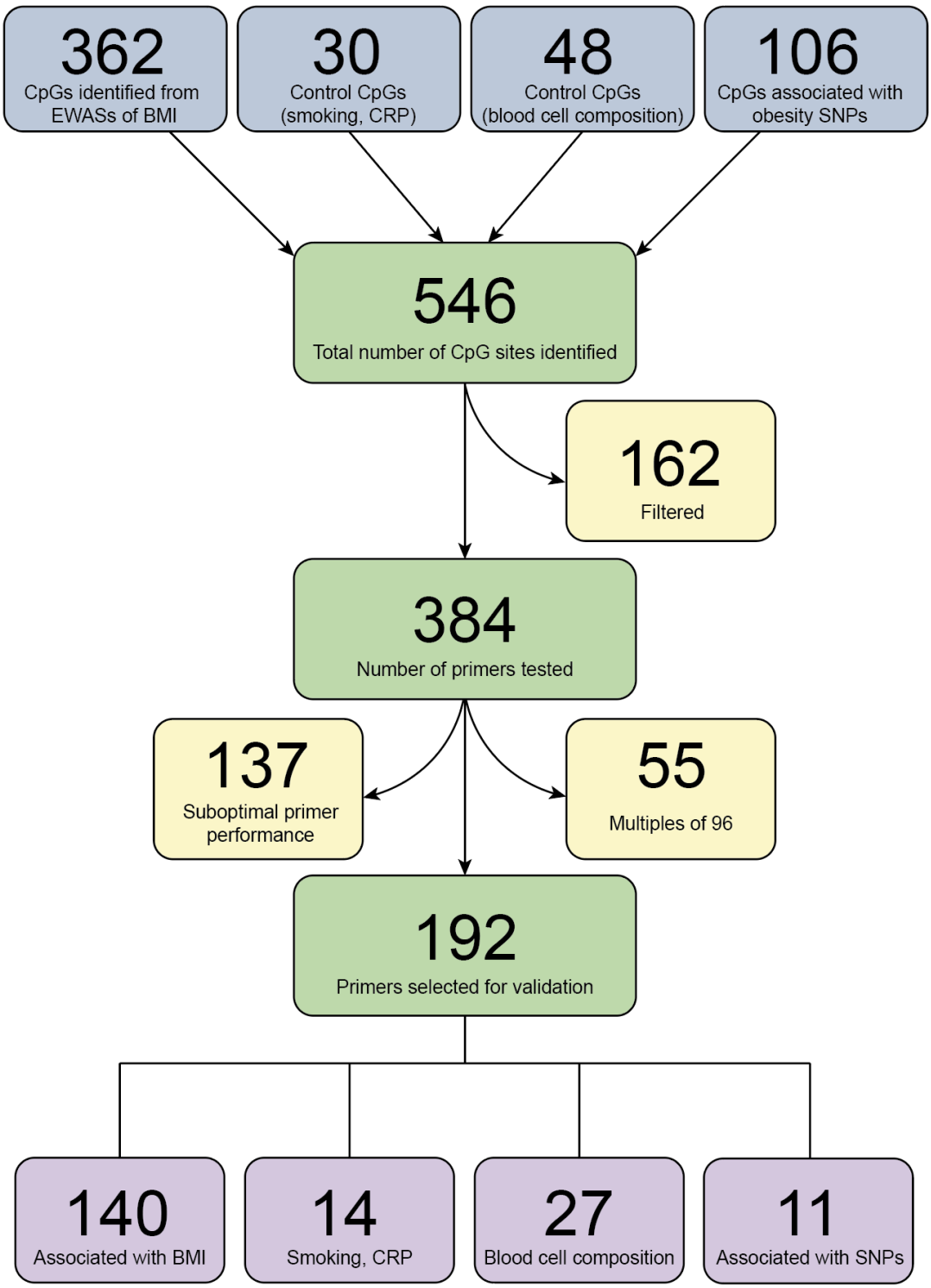


Figure 2-11. Pipeline for selecting obesity associated CpG sites for validation.

384 primer pairs (forward and reverse) were generated in silico using Primer3 with a customised script (215). Primers were designed to uniquely align to their target sequence and have an annealing temperature of 55°C. An additional sequence was added to the 5' end of the forward and reverse primers for compatibility with downstream use of the Fluidigm Access Array microfluidics platform 2.15.

Forward primer tag: 5' ACACTGACGACATGGTTCTACA 3'

Reverse primer tag: 5' TACGGTAGCAGAGACTTGGTCTA 3'

The primers were purchased from Sigma-Aldrich (Sigma-Aldrich, Gillingham, Dorset).

2.14.1 Evaluation of primers

The 384 primers were objectively evaluated using a classification system to qualitatively test their efficacy and specificity. This was to ensure optimal function of primers when used in the microfluidics based assay, and thus was designed to as closely as possible mimic the conditions used in the next steps of the experiment. The input genomic DNA for primer evaluation was extracted from blood or CD4+ T-lymphocytes using the same protocol as described in section 2.8.1. The samples were, however, from different individuals to those included in the study of obesity associated CpG methylation. Bisulfite conversion was performed on 500ng genomic DNA as described in section 2.11.

The evaluation of primers was performed by Dr Michelle Holland and Adrian Signell at Kings College London (KCL), and will only be discussed briefly below. Each of the primers were amplified using PCR in a reaction consisting of 2.5-5 ng bisulfite converted DNA (or the appropriate volume of ultrapure water for negative controls), Taq polymerase, dNTPs, $MgCl_2$, appropriate buffer as well as the forward and reverse primers. The resulting amplicons were separated by size using agarose gel electrophoresis. Primers were evaluated using two sets of PCR agents from different manufacturers to determine which system would be used to generate sequencing libraries in subsequent experiments. Thus, the Roche FastStart High Fidelity PCR

System was compared to the Qiagen HotStarTaq DNA Polymerase (216, 217). Following gel electrophoresis, gels were imaged, and a scoring system was used to indicate the specificity and efficiency of each primer.

Specificity of a primer was objectively measured by comparing the size of the observed amplicon to the expected size, whereas efficiency of a primer was objectively measured by comparing the relative brightness of the observed amplicon to a positive control, and took into consideration the contrast between the observed amplicon and background. The scoring system ranged from 1-5, where 1 indicated poor primer performance in terms of specificity and efficiency, and 5 indicated optimal primer performance. Further, the presence of primer dimers (usually ~100bp in size) was noted, but did not influence primer scoring. Primer dimers will be instead be removed through a size selection step, in which products under 150 bp will be removed prior to library preparation to eliminate primer contamination in the subsequent sequencing. Primers classified with a score of 3 or above were considered successful and compatible with the experimental protocol to be used with the Fluidigm Access Array microfluidics platform in the next phase of the study (section 2.15).

Overall, it was found that the Qiagen PCR System resulted in primers with higher efficiency and specificity using the given experimental conditions. Indeed, with the Qiagen system, 247 out of the 384 primers tested scored 3 or higher, whereas only 207 out of 385 of the primers scored 3 or higher when using the Roche system. It was therefore decided that the Qiagen system would be used in future experiments (in the larger cohort of 1000 lean versus 1000 obese males). Primers selected for the Bis-PCR-Seq were therefore selected among the 247 relatively highly scoring primers.

As the next steps of the protocol run in multiples of 96, 192 out of the 247 relatively highly scoring primers were selected for validation in the cohort of 96 lean versus 96 obese males. A small number of primers (n = 19) that scored were included despite scoring <3 but were of special interest as they had been replicated in at least two

EWASs of BMI. The complete list of the 192 primers and their scores in the evaluation assay can be found in Appendix 3.

2.15 The Fluidigm Access Array

The Fluidigm Access Array is a multiplex bisulphite PCR sequencing microfluidics-based assay that allows analysis of 10s to 100s of targeted CpG sites simultaneously. This assay was used to validate the 192 selected obesity associated CpG sites in DNA from 96 lean and 96 obese males. Thus, following bisulfite conversion of the sample DNA, PCR reactions were performed using the Fluidigm Access Array, in which several target specific primers are amplified in parallel for each DNA sample. The products of the multiple PCR reactions were harvested from the Fluidigm Access Array Integrated Fluidic Circuit (IFC) for downstream sequencing (218, 219). The IFC process, library preparation and subsequent NGS was performed at the Genome Centre Facility at Charterhouse Square, QMUL, using the 48.48 layout on the Fluidigm® C1 system (Fluidigm® USA).

The Access Array System library preparation workflow is summarised in Figure 2-12.

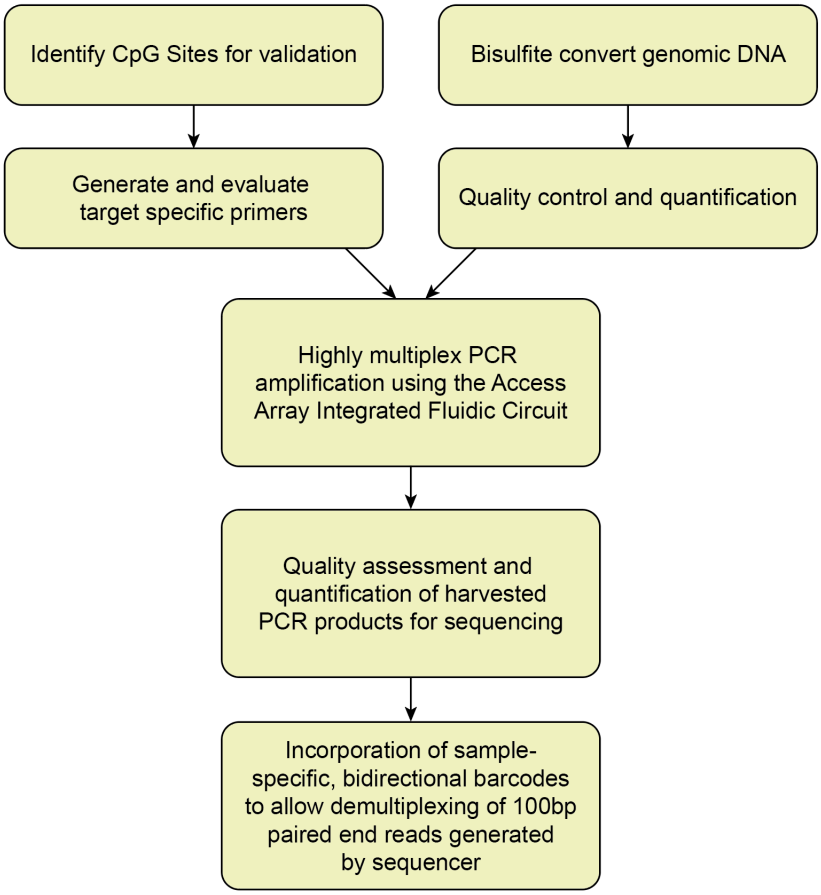


Figure 2-12. The Access Array System library preparation workflow.
bp = base pair

2.15.1 Preparation of primers

192 primers were selected for validation as described in section 2.14. In order to comply with the 48.48 layout on the Fluidigm® C1 system (Fluidigm® USA), four primers (four forward and four reverse) were combined in each well. The primer concentration was 2.5 uM for each of the forward and reverse primers diluted in a total of 100 uL ultrapure water. The primers were pipetted into the IFC primer inlets (Figure 2-13).

2.15.2 Preparation of DNA samples

500ng of blood-derived DNA from the 192 samples (from 96 lean and 96 obese males) was bisulfite converted as described in section 2.11. Bisulfite converted DNA was diluted to a concentration of 11 ng/µL using a Qubit 3.0 Fluorometer (Life Technologies). 45 µL of DNA (i.e. approximately 500ng) was used from each participant. DNA samples were randomised onto two 96-well plates before being pipetted into the sample inlets on the Fluidigm C1 integrated fluidic circuit array chip (Figure 2-13).

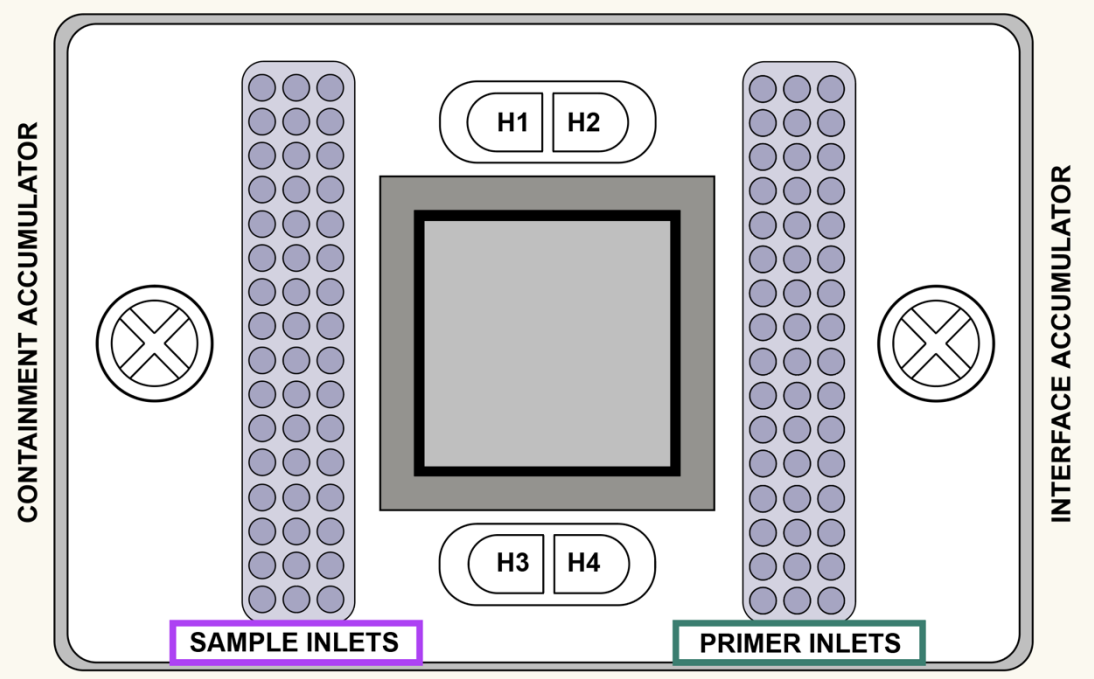


Figure 2-13. The Fluidigm C1 integrated fluidic circuit array chip
Bisulfite converted DNA samples and barcode oligonucleotides were loaded into the sample inlets. The locus-specific primers were loaded into the primer inlets. Pneumatic pressure then drives the DNA samples, barcode oligonucleotides and primers into the chip's ultra-tiny microfluidic reaction chambers (represented as the middle grey area in the image above).
H1 = H1 well, H2 = H2 well, H3 = H3 well, H4 = H4 well.

2.15.3 Multiplex DNA methylation analysis of target regions

An overview of the protocol for multiplex DNA methylation analysis of target regions is presented in Figure 2-14 (219).

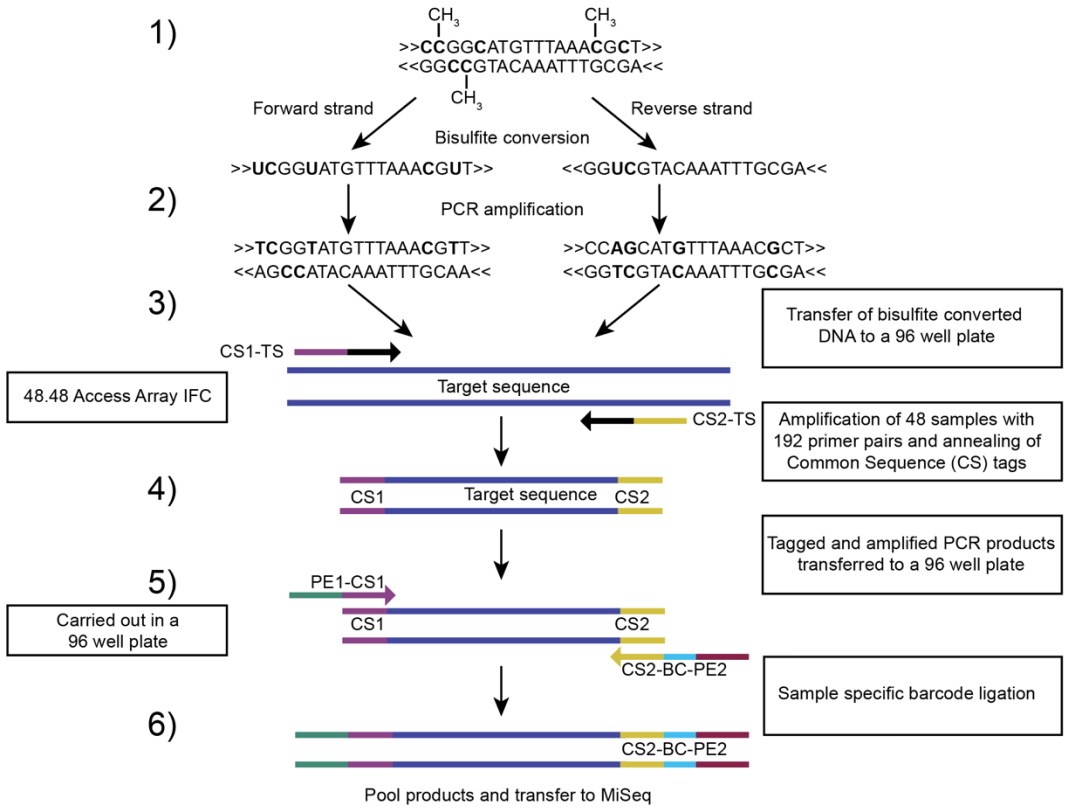


Figure 2-14. Overview of the protocol for multiplex DNA methylation analysis of target regions. CS1 = Common Sequence tag 1, CS2 = Common Sequence tag 2, TS = Target-Specific primer Sequence, PE1 = paired end sequence 1, PE2 = Paired End sequence 2, BC = Barcode Sequence, IFC Integrated Fluidic Circuit. Adapted from (219).

With refence to the steps in Figure 2-14, the process involves the following:

1) Genomic DNA is bisulfite converted; unmethylated cytosines into uracils while leaving methylated cytosines unchanged
2) During PCR amplification, uracil is amplified as thymine
3) Multiplex PCR in the Fluidigm Access Array IFC. During this PCR process, the common sequence (CS) tags annealed to the target specific primer pairs are amplified along with the target specific primer sequences. The common sequence tag for the forward primer is 5' ACACTGACGACATGGTTCTACA 3'

and the common sequence tag for the reverse primer is 5' TACGGTAGCAGAGACTTGGTCTA 3'.

4) The tagged and amplified PCR products are transferred to a 96 well plate
5) PCR products are ligated to sample-specific barcodes. This means that PCR products can be pooled for sequencing. Paired-end sequences are also ligated to comply with subsequent sequencing using the Illumina MiSeq sequencing system
6) PCR products are pooled, harvested and transferred for sequencing

2.15.4 Sequencing of PCR products

Sequencing of the pooled amplicons was performed using the Illumina MiSeq sequencing system (Illumina, Sand Diego, California) (220). The V2 reagent kit was used. The MiSeq sequencing system employs NGS and is thus a high throughput sequencing technology based on the principle of running millions of amplified DNA sequences in parallel. It is particularly adapted for targeted resequencing. In addition to providing information on CpG methylation levels at the particular site selected in each primer, the method provides cis genetic information at single nucleotide resolution in the area surrounding the CpG site (220).

The first step in sequencing of the pooled PCR products is cluster amplification. First, the library was loaded into a flow cell. The illumina patterned flow cell is an 8-channel sealed glass micro fabricated device on which surface are millions of nanowells covered with short oligonucleotide sequences (seeding primers) (221). After flow cell assembly, the PCR products are deposited into the nanowells. During cluster generation, only one single DNA template is able to hybridise and form a cluster within each nanowell. Each bound fragment is amplified as a clonal cluster through bridge amplification. Thus, a monoclonal cluster is formed in each nanowell consisting of approximately 1 million copies of the template DNA (221).

The next step sequencing-by-synthesis. DNA polymerase is added together with all four nucleotides, each carrying a base-unique fluorescent label and with the 3'-OH group chemically blocked such that each incorporation becomes a unique event.

After incorporation, an imaging step occurs prior to the 3’ blocking group being removed to prepare the strand for the next incorporation. The emission wavelength and intensity are used to identify the base. Paired-end sequencing was used to increase the number of reads per sequence and improve accuracy when later aligning the fragments to a reference genome.

Following sequencing-by-synthesis, the reads were demultiplexed by using the sample specific barcode sequences before being aligned to an in silico bisulfite converted reference genome. The final NGS data output was uploaded as binary files for each sample onto BaseSpace.

Variant calling was used to identify the percentage of 5mC at each CpG site; unmethylated CpG sites in the target sequences would have T whereas methylated CpG sites in the target sequences would have C in the mapped reads.

Chapter 3
The Impact of Paternal Metabolic Health on Offspring Birth Weight

3.1 Introduction

Infants born with low birth weight are predisposed to developing diabetes and cardiovascular disease in later life (60). The most common cause of fetal growth restriction is poor placental development and invasion, but the reason for this deficiency is usually unclear. Other factors include maternal disease, maternal toxins such as cigarette smoke and, in less economically developed settings, suboptimal maternal nutrition (222). Most cases of fetal growth restriction remain idiopathic and it can be difficult to distinguish between infants affected by fetal growth restriction caused by a failure of a fetus to meet its growth potential and infants who are constitutionally small (222). It is therefore of considerable clinical importance to improve our ability to understand the risk factors, pathophysiology and prevention of fetal growth restriction.

Animal studies suggest that paternal metabolic syndrome is associated with low offspring birth weight, as well as with an increased risk of metabolic disease in adulthood, when controlling for maternal factors (138, 223).

In humans, whereas maternal obesity and insulin resistance predisposes her offspring to be large for gestational age (LGA), paternal obesity and insulin resistance predispose his offspring to be small for gestational age (SGA) (68-70, 224-226). Rare genetic variants that cause monogenic diabetes in fathers are associated with both low birth weight of his offspring and an increased risk of offspring diabetes later in life (124). It is possible that less penetrant but more common genetic or epigenetic variants that are associated with insulin resistance in fathers, are also associated with reduced birth weight of his offspring.

The association between paternal obesity and reduced offspring birth weight has not been replicated across studies (227). Further, studies in this area have mostly been retrospective, and sometimes relied on surveys that ask mothers to convey the weight/BMI of the father, potentially reducing the reliability of results. I therefore undertook a prospective cohort study, 'The Dad's Health Study', in which I performed

detailed phenotypic measurements of fathers early in their partners' pregnancies, focussing in particular on metabolic disease risk factors. In order to reduce confounding from maternal factors, I excluded mothers who smoked or who had chronic disease. I followed up pregnancies and noted pregnancy outcomes. I collected peripheral blood from the mothers and umbilical cord and childbirth. I used customised growth centiles rather than raw offspring birth weights in all analyses. This involved adjusting for maternal BMI, ethnicity, parity, length of gestation at delivery and neonatal sex.

My aim was to identify paternal metabolic risk factors for poor intrauterine growth of his offspring. Identifying such factors would inform public health policies directed towards improving paternal health before conception and potentially improving the health of the next generation.

3.1.1 Declarations

Recruitment, sample collection, processing and DNA extraction for the Dad's Health Study was performed by myself and a research midwife (Anna Greco). Statistical analyses were carried out by myself under guidance from a senior IT Trainer at UCL (Jim Tyson), and Professor Aviva Petrie at the UCL Eastman Dental Institute.

3.2 Hypothesis

Paternal obesity and/or insulin resistance increases the risk of fathering small for gestational age (SGA) offspring.

3.3 Specific Objectives

In a prospective cohort study:

1. To identify whether paternal obesity, as measured by BMI, is associated with an increased risk of fathering SGA offspring.
2. To identify whether paternal insulin resistance, as measured by HOMA-IR, is associated with an increased risk of fathering SGA offspring.

3.4 Methods

Methods for the prospective cohort study, hereafter referred to as 'The Dad's Health Study', are described in detail in Chapter 2 section 2.3-2.8.

The study was carried out at University College London Hospital between May 2016 and June 2019. Favourable ethical approval for the study was granted from the South East Coast - Surrey Research Ethics Committee on 28 September 2015 (REC reference number 15/LO/1437, IRAS project ID 164459). The study was also registered with the UCLH Joint Research Office (Project ID 15/0548). All participants provided written, informed consent.

3.4.1 Study Design and Population

Recruitment for the Dad's Health Study is described in detail in Chapter 2 section 2.6-2.8.

Briefly, the Dad's Health Study aimed to investigate the association between paternal metabolic health and the risk of fathering a small for gestational age infant, here defined as an infant with a birth weight lower than the 10th customised birth weight centile (176).

Paternal obesity was hypothesised to either be associated with an overall reduction in mean offspring birth weight, or with an increase in the proportion of offspring affected by FGR. Therefore, two different approaches to determining an adequate sample size were employed (discussed in detail in Chapter 2 section 2.3.4). In the first scenario, a sample size of 151 obese and 151 lean fathers was calculated to be required to achieve 90% power to detect a difference in offspring birth weight of 150 grams (at 5% significance). In the second scenario, it was calculated that 219 lean and 219 obese fathers would be required to detect a 5% difference in the risk of fathering FGR offspring with 80% power (at 5% significance). It was decided to recruit a total of 500 fathers and their partners to allow for participant drop-out.

Infants were classified as SGA if their customised birth weight centile was <10, classified as AGA if their customised birth weight centile 10-90 and classified as LGA if their customised birth weight centile was ≥ 90 using the Perinatal Institute's customised birth weight centile charts (176, 228).

3.4.2 Paternal Phenotype Measurements

Paternal phenotype measurements are described in detail in Chapter 2 section 2.3-2.8.

Briefly, fathers were recruited antenatally, typically around the time of their female partner's (the mothers) first ultrasound scan (10-14 weeks of gestation). The fathers were asked to be fasted for 8-10 hours prior to attending for a study visit at UCLH. Their phenotype was measured by a trained research doctor or midwife. This included measures of their height, weight, blood pressure and waist circumference. Peripheral blood was taken from the median cubital vein and sent for measures of insulin, glucose, haemoglobin, c-reactive protein (CRP), haemoglobin A1c (HbA1c), lipid levels, liver function and kidney function. Insulin resistance was calculated using HOMA as detailed in Chapter 2 section 2.7. Fathers were asked to fill in a questionnaire enquiring about past medical, family and treatment history (Appendix 1).

Two purple (EDTA-coated) and one gold top vacutainers were ultracentrifuged for 15 minutes at 1000g within one hour of venepuncture. Buffy coat from the EDTA-coated vacutainer was used for DNA extraction as described in Chapter 2 section 2.8.

As part of an additional study to investigate a potential genetic or epigenetic mechanism linking paternal metabolic health with offspring birth weight, participating fathers were also asked if they would provide a semen sample. Approximately one third (181/500) of participating fathers consented to providing a semen sample. Semen samples were analysed and processed as described in Chapter 2 section 2.10. Results from DNA methylation analyses of semen samples are presented in Chapter 5.

3.4.3 Maternal Phenotype Measurements

Female participants, the mothers, were typically seen by a member of the research team in conjunction with their 26 to 28-week Glucose Challenge Test (GCT).

The mothers completed a questionnaire enquiring about obstetric, medical, family and treatment history (Appendix 2). Information regarding maternal ethnicity, height, weight, obstetric history and estimated date of delivery was obtained from electronic antenatal records. Clinical biochemistry information, including results from the glucose challenge test and, where relevant, the glucose tolerance test, were also obtained from electronic antenatal records. A peripheral blood sample was obtained from the median cubital vein. Two purple (EDTA-coated) and one gold top vacutainers were ultracentrifuged for 15 minutes at 1000g within one hour of venepuncture. Buffy coat from the EDTA-coated vacutainer was used for DNA extraction as described in Chapter 2 section 2.8.

3.4.4 Offspring Sample Collection and Measurements

At childbirth, umbilical cord blood samples were obtained shortly following the delivery of the placenta, either by myself or the attending midwife. A 10 mL sample was obtained from the umbilical arteries or vein. Samples were ultracentrifuged for 15 minutes at 1000 g and DNA was extracted from buffy coat as described in Chapter 2 section 2.8.

Detailed information about the pregnancy and delivery was recorded, including any complication during the pregnancy, length of gestation, mode of delivery, any anaesthesia used, offspring Apgar scores at one and five minutes following birth as well as any neonatal complications. An Apgar score, named after the obstetric anaesthetist Dr Virginia Apgar, is a score developed to rapidly assess the physical condition of a newborn infant and the need for prompt intervention to establish breathing (229). The score comprises five components which also correspond to the letters in the name Apgar; ‘Appearance, Pulse, Grimace, Activity, and Respiration’ (229).

3.4.5 Statistical Analyses

Phenotype analyses were carried out using RStudio version 1.1.456. The cut-off value for significance was taken as 0.05 in all analyses unless otherwise stated.

Measurements of BMI, waist circumference, systolic and diastolic blood pressure, total cholesterol as well as HDL and LDL cholesterol were found to be approximately normally distributed (Shapiro-Wilk test $p>0.10$) among study participants. For these measurements, results are summarised as mean values and standard deviations. The means of these measurements were compared between the discovery, replication and obesity cohorts using a one-way analysis of variance (ANOVA). This was followed by Tukey's Honest Significant Difference test if $p < 0.05$. The distribution of measurements of fasting glucose, C-reactive protein and triglyceride levels as well as calculated HOMA-IR values were found to be skewed (Shapiro-Wilk test $p<0.10$). These are therefore summarised as median values and interquartile ranges. In these cases, the Kruskal-Wallis test of ranks was used to compare median values between the cohorts. This was followed by Dunn's test of multiple comparisons using rank sums if $p < 0.05$.

3.5 Results

3.5.1 Recruitment Numbers to Study Groups

As discussed in Chapter 2 section 2.3.4, the recruitment target was set to 500 participating couples (mothers and fathers) in order to allow for participant dropout for reasons such as miscarriage, multiple pregnancy that was unknown at the time of recruitment and transfer of care to other hospitals. Recruitment was therefore capped after 500 couples had been recruited. A total of 110 participants were excluded from analyses of the paternal impact on fetal growth for reasons detailed in Figure 3-1. Thus, a total number of 390 couples were studied with regards to the study objectives.

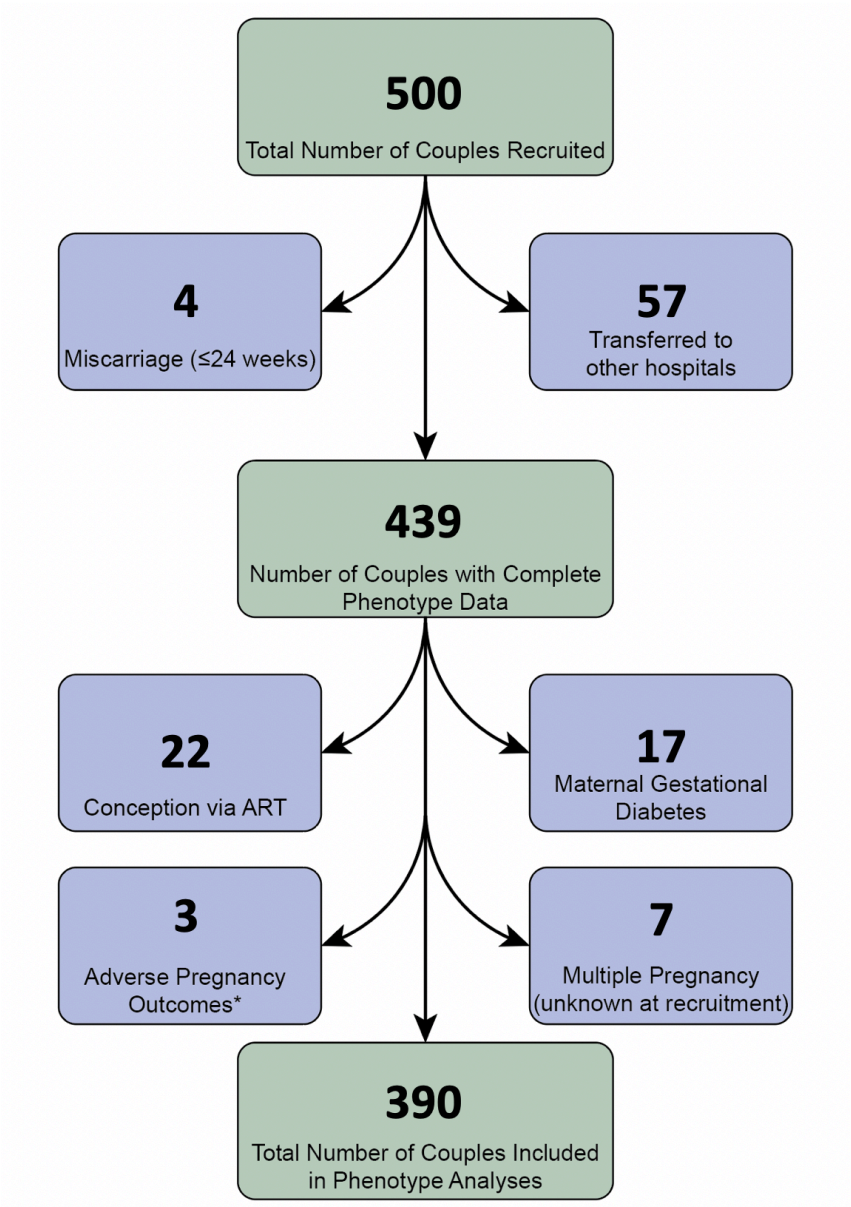


Figure 3-1. Flow-chart of participants included in analyses of the paternal impact on fetal growth in the Dad's Health Study, and reasons for participant exclusion.
Participants were mainly excluded from final analyses on the basis of factors that are known to impact on fetal growth but were unknown at the time of recruitment the Dad's Health study. These include maternal gestational diabetes, significant maternal disease and multiple pregnancy. Couples were also excluded if their antenatal care was transferred to another hospital such that pregnancy and delivery details could not be recorded, or in cases of miscarriage. Pregnancies resulting from ART were also excluded.
*Adverse pregnancy outcomes refers to two cases of fetal death (>24 weeks of pregnancy) and one case of maternal disease in pregnancy that led to the delivery of her offspring at 24 weeks of gestation.
ART = Assisted Reproductive Technologies.

	Excluded	Included	p
n	49	390	
Age (years). Mean (SD)	38.1 (5.3)	36.3 (5.2)	0.019
BMI (kg/m2). Mean (SD)	26.8 (3.6)	26.3 (4.1)	0.426
Waist circumference (cm). Mean (SD)	93.9 (9.8)	91.9 (12)	0.249
SPB (mmHg), average of two measurements. Mean (SD)	119 (16)	123 (13)	0.065
DPB (mmHg), average of two measurements. Mean (SD)	76 (11)	79 (9)	0.022
Total cholesterol (mmol/L). Mean (SD)	4.9 (0.9)	5.0 (0.9)	0.523
HDL cholesterol (mmol/L). Mean (SD)	1.4 (0.3)	1.5 (0.4)	0.182
LDL cholesterol (mmol/L). Mean (SD)	3.0 (0.8)	3.0 (0.8)	0.667
Fasting glucose (mmol/L). Median (IQR)	4.9 (0.5)	4.8 (0.5)	0.034
Fasting insulin (mIU/L). Median (IQR)	8.0 (6.4)	6.6 (5.5)	0.060
HOMA-IR. Median (IQR)	1.8 (1.4)	1.4 (1.2)	0.038
HOMA2-IR. Median (IQR)	1.0 (0.8)	0.9 (0.7)	0.056
CRP (mg/L). Median (IQR)	0.9 (0.7)	0.8 (0.8)	0.528
Triglycerides (mmol/L). Median (IQR)	1.1 (0.4)	1.0 (0.7)	0.958

Table 3-1. Phenotype comparisons between males who were excluded from the study after their study visit and males who were included in analyses of the paternal impact of fetal growth.
There were small significant metabolic differences between males who initially partook in the Dad's Health Study but were thereafter excluded for reasons such as transfer of antenatal care to another hospital or miscarriage. These differences were in diastolic blood pressure, fasting glucose and HOMA-IR. Fathers who were excluded were also significantly older.

There were small but significant differences between excluded and included fathers in diastolic blood pressure, fasting glucose and HOMA-IR (Table 3-1). Fathers who were excluded were also significantly older. Most fathers were excluded because their partner developed gestational diabetes, or that the couple had conceived via ART. It is possible that men whose female partners developed diabetes were also more likely to be insulin resistant themselves ('assortative mating'). It is also possible that couples who undergo ART are on average older than couples who conceive naturally. These would be potential explanations for the differences between excluded and included participating fathers.

3.5.2 Paternal Phenotype Profiles

Phenotype characteristics of male participants (the fathers) are presented in Table 3-2.

There were clear, significant differences in metabolic characteristics between the three groups of male participants (lean, overweight and obese). Thus, measurements of BMI, waist circumference, systolic and diastolic blood pressure, HDL cholesterol, fasting glucose and insulin, HOMA-IR, HOMA2-IR, CRP and triglycerides were all significantly different at the $p = < 0.05$ level between the three groups (for details of significance see Table 3-2). In addition, there were significant differences in total and LDL cholesterol between the lean and overweight groups and between the lean and obese groups of male participants. The difference in BMI between the three groups of participating males is visualised in Figure 3-2.

It should be noted that there was a slight but significant difference in age between the lean and obese groups of participating fathers ($p = 0.025$). This was included as a potential confounding factor in later analyses of the association between paternal metabolic health and offspring birth weight.

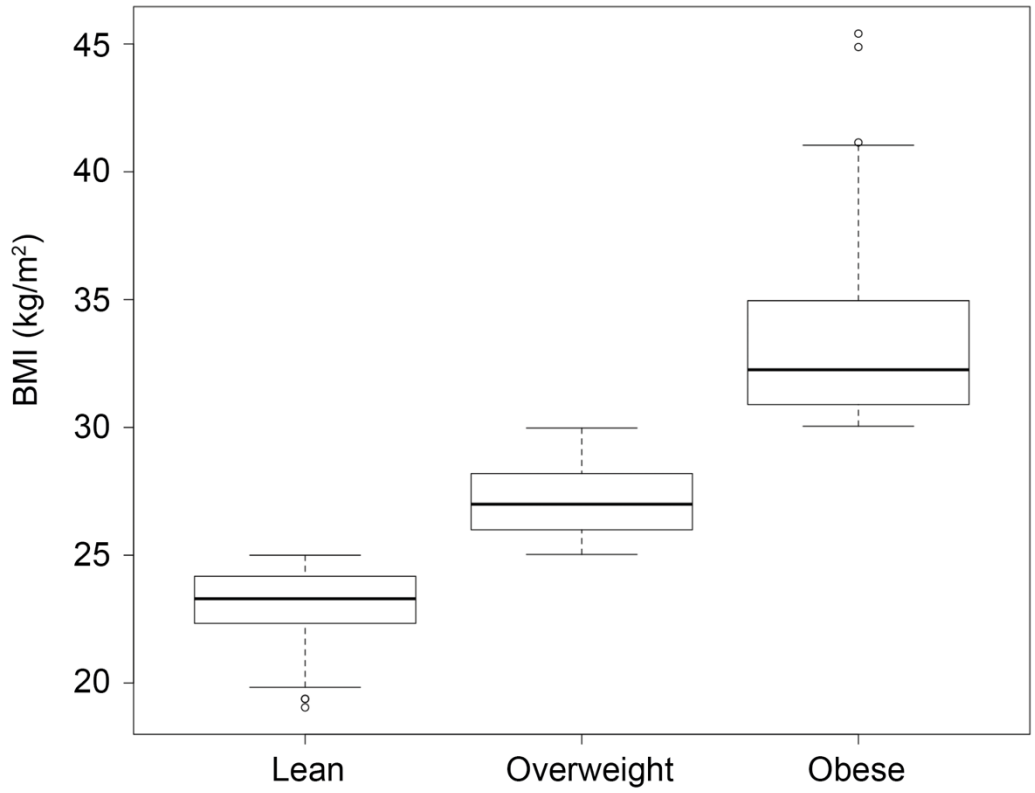


Figure 3-2. Comparison of BMI (kg/m^2) between participants in the groups of lean, overweight and obese participants respectively.
The boxes indicate the middle 50% of values, with the strong black line representing the median values. Outliers are indicated by empty circles.

In summary, three groups of metabolically distinct male participants and their female partners were recruited and followed up with regards to pregnancy outcome.

	Lean	Overweight	Obese	Healthy Reference Range	p (lean versus overweight)	p (lean versus obese)	p (overweight versus obese)
n	179	147	64				
Age (years). Mean (SD)	35.8 (4.8)	36.2 (4.6)	37.7 (6.6)		0.727	0.025	0.122
BMI (kg/m^2). Mean (SD)	23.1 (1.3)	27.1 (1.4)	33.5 (3.6)	18.5–24.9	<0.001	<0.001	<0.001
Waist circumference (cm). Mean (SD)	83.4 (6)	94.5 (6)	110 (6)	<94cm	<0.001	<0.001	<0.001
SPB (mmHg), average of two measurements. Mean (SD)	118 (13)	123 (10)	134 (11)	90 - 120	<0.001	<0.001	<0.001
DPB (mmHg), average of two measurements. Mean (SD)	76 (8)	80 (8)	86 (9)	60 - 80	<0.001	<0.001	<0.001
Total cholesterol (mmol/L). Mean (SD)	4.9 (0.9)	5.1 (0.9)	5.3 (0.9)	< 5.0	0.029	<0.001	0.430
HDL cholesterol (mmol/L). Mean (SD)	1.6 (0.3)	1.5 (0.4)	1.2 (0.3)	0.9-1.5 (males)	<0.001	<0.001	<0.001
LDL cholesterol (mmol/L). Mean (SD)	2.8 (0.8)	3.1 (0.8)	3.1 (0.7)	< 3.5	0.009	0.036	0.970
Fasting glucose (mmol/L). Median (IQR)	4.7 (0.5)	4.8 (0.6)	5.0 (0.5)	3.9-5.8	0.041	<0.001	0.017
Fasting insulin (mIU/L). Median (IQR)	5.2 (3.2)	7.2 (5.1)	13 (11)	2.6-24.9	<0.001	<0.001	<0.001
HOMA-IR. Median (IQR)	1.1 (0.7)	1.6 (1.3)	2.8 (2.6)	Variable, often ≤ 2.0	<0.001	<0.001	<0.001
HOMA2-IR. Median (IQR)	0.7 (0.4)	0.9 (0.7)	1.6 (1.5)		<0.001	<0.001	<0.001
CRP (mg/L). Median (IQR)	0.6 (0.3)	0.9 (1.0)	1.5 (2.4)	0-5.0	<0.001	<0.001	<0.001
Triglycerides (mmol/L). Median (IQR)	0.8 (0.4)	1.2 (0.7)	1.7 (1.4)	<2.3	<0.001	<0.001	<0.001

Table 3-2. Phenotypes of Male Participants in the Dad's Health Study
Reference ranges are derived from the UCLH Clinical Biochemistry Test Information sheet available from (230). The reference range for HOMA-IR is derived from (231). The HOMA2-IR reference range is derived from (232). The reference range for waist circumference is derived from (179). The reference ranges for blood pressure are derived from (233). SD = Standard Deviation, IQR = interquartile range, BMI = Body Mass Index, SBP = Systolic Blood Pressure, DBP = Diastolic Blood Pressure, HOMA-IR = Homeostatic Model Assessment of Insulin Resistance, CRP = C-Reactive Protein, HDL = High Density Lipoprotein, LDL = Low Density Lipoprotein.

3.5.3 Maternal Phenotype Profiles

Phenotype characteristics of female participants (the mothers) are presented in Table 3-3.

There was a positive relationship between maternal and paternal BMI (R^2 = 0.061, p = < 0.001, Figure 3-2), such that overweight and obese men were more likely to have overweight and obese partners (Table 3-3).

Ideally, an investigation of the impact of paternal phenotype on fetal growth should control for maternal phenotype, as the direct 'environmental effect' of the mother during pregnancy is likely to overshadow a potential genetic or epigenetic contribution from the father. However, I adjusted for maternal BMI by using customised birth weight centiles (which take maternal height and weight into account) rather than raw birth weight (176). Thus, it was felt that despite there being a positive association between parental BMI measurements, birth weight centiles could still accurately be compared between lean, overweight and obese fathers.

	Female partners of lean males	Female partners of overweight males	Female partners of obese males	Reference Range	p-value (difference between groups)	p-value (lean versus over-weight)	p-value (lean vs obese)	p-value (over-weight vs obese)
n	179	147	64					
Age (years). Mean (SD)	34.2 (4.4)	34.6 (3.7)	34.2 (6.3)	N/A	0.648			
BMI (kg/m2). Mean (SD)	22.8 (3.4)	24.0 (3.9)	25.3 (5.2)	18.5–24.9	<0.001	0.012	<0.001	0.087
Booking glucose (mmol/L). Median (IQR)	4.4 (0.7)	4.4 (0.6)	4.4 (0.7)	<6.1	0.970			
GCT result (mmol/L). Median (IQR)	6.0 (1.8)	6.1 (1.7)	6.2 (1.6)	<7.8	0.542			
Parity. Median (IQR)	0 (0)	0 (0)	0 (1)	N/A	0.062			

Table 3-3. Comparison of the metabolic phenotype of female partners of lean, overweight and obese males (fathers) respectively.
The 'booking glucose' refers to a non-fasted blood glucose test that pregnant women are offered at their first antenatal appointment (typically 5-9 weeks of gestation) and provides an indication of whether further investigation of diabetes is warranted. The reference range for the GCT is taken from (178). BMI = Body Mass Index, GCT = Glucose Challenge Test, SD = Standard Deviation.

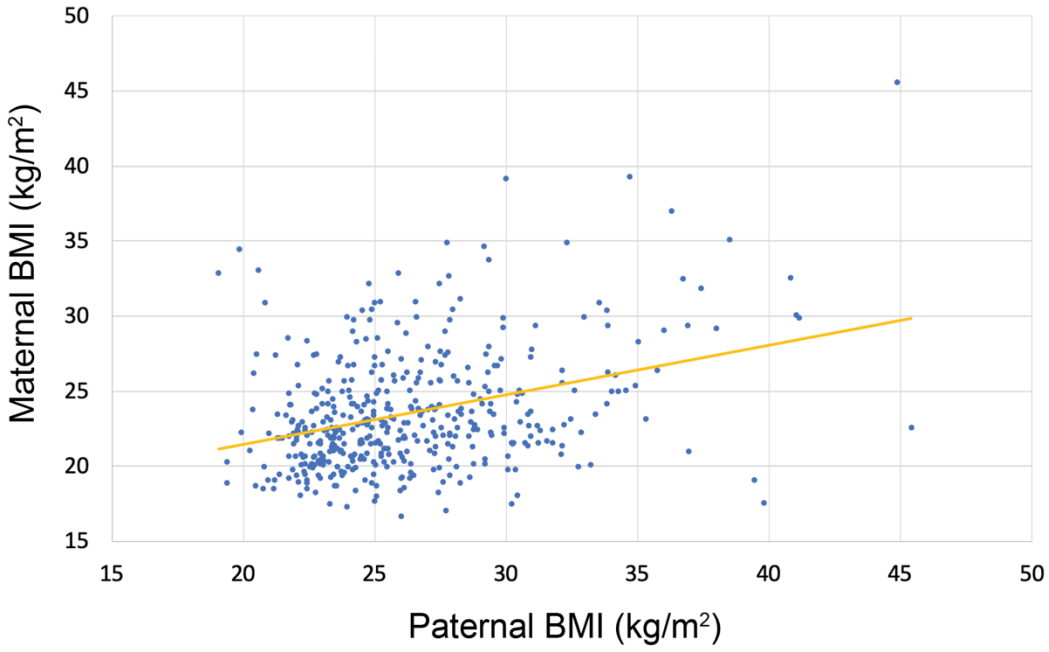


Figure 3-3. Comparison of BMI measurements for male and female participants
There was a positive relationship between maternal and paternal BMI (R^2 = 0.061, p = < 0.001). This graph shows the influence of assortative mating.

3.5.4 Offspring Phenotype Profiles

Phenotypes of the 390 infants are presented in Table 3-4.

	Offspring of lean males	Offspring of overweight males	Offspring of obese males	Reference Range	P (difference between groups)
n	179	147	64		
Customised centile. Mean (SD)	44.5 (27.6)	43.3 (27.3)	46.7 (26.6)	N/A	0.715
Length of gestation (days). Mean (SD)	280 (10)	280 (8.5)	280 (9.4)	259 – 294	0.928
Apgar score at 1 minute. Median (ICR)	9 (0)	9 (1)	9 (0)	>7	0.061
Apgar score at 5 minutes. Median (ICR)	10 (0)	10 (0)	10 (0)	>7	0.452

Table 3-4. Comparison of offspring phenotypes between groups of lean, overweight and obese fathers
The reference range for Apgar scores is derived from (15). Customised centiles are calculated using (9). SD = Standard Deviation.

Contrary to my hypothesis, I found that lean, overweight and obese fathers had offspring with similar mean customised birth weight centiles.

3.5.5 Offspring Birth Weight Centiles in Relation to Paternal Metabolic Health

At the time of study completion, a total number of 48 SGA, 324 AGA and 18 LGA infants had been born to parents participating in the Dad's Health Study. The mean customised birth weight centiles for these infants were 4.6 (SD 3.1), 47.6 (SD 22.4) and 95.6 (SD 3.5) respectively. The paternal metabolic profiles of SGA, AGA and LGA infants is presented in Table 3-5.

	Fathers of SGA offspring	Fathers of AGA offspring	Fathers of LGA offspring	Healthy Reference Range	P (difference between groups)
n	48	324	18		
Age (years). Mean (SD)	35.5 (4.5)	36.2 (5.2)	37.9 (3.9)		0.215
BMI (kg/m2). Mean (SD)	26.5 (4.3)	26.1 (4)	27.4 (4)	18.5–24.9	0.411
Waist circumference (cm). Mean (SD)	92.9 (12.2)	91.3 (11.2)	96.4 (12.9)	<94cm	0.137
SPB (mmHg). Mean (SD)	122 (14)	123 (12)	125 (12)	90 - 120	0.655
DPB (mmHg). Mean (SD)	80 (9)	79 (9)	83 (12)	60 - 80	0.149
Total cholesterol (mmol/L). Mean (SD)	5.0 (0.8)	5 (0.9)	5 (0.8)	< 5.0	0.992
HDL cholesterol (mmol/L). Mean (SD)	1.5 (0.4)	1.5 (0.4)	1.5 (0.3)	0.9-1.5 (males)	0.935
LDL cholesterol (mmol/L). Mean (SD)	3.0 (0.7)	3 (0.8)	3 (0.7)	< 3.5	0.982
Fasting glucose (mmol/L). Median (IQR)	4.9 (0.6)	4.8 (0.5)	4.7 (0.2)	3.9-5.8	0.787
Fasting insulin (mIU/L). Median (IQR)	7.3 (4)	6.4 (5.5)	6.5 (12)	2.6-24.9	0.444
HOMA-IR. Median (IQR)	1.6 (1.0)	1.4 (1.2)	1.5 (2.6)	Variable, often ≤ 2.0	0.395
HOMA2-IR. Median (IQR)	0.9 (0.6)	0.8 (0.7)	0.9 (1.6)	Variable, often ≤ 1.7	0.456
CRP (mg/L). Median (IQR)	1.0 (1.5)	0.7 (0.7)	0.8 (1.1)	0-5.0	0.081
Triglycerides (mmol/L). Median (IQR)	1.0 (0.7)	1.1 (0.7)	0.9 (0.9)	<2.3	0.833

Table 3-5. Phenotype characteristics of fathers of SGA, AGA and LGA offspring
SGA = Small for Gestational Age, AGA = Appropriate for Gestational Age, LGA = Large for Gestational Age, BMI = Body Mass Index, SBP = Systolic Blood Pressure, DBP = Diastolic Blood Pressure, HOMA-IR Homeostatic Model Assessment of Insulin Resistance.

As presented in Table 3-5, this study did not identify any paternal metabolic trait that was significantly associated with the risk of fathering SGA offspring. There was, however, a trend towards higher insulin resistance in fathers of SGA offspring when compared to AGA offspring.

The relationship between offspring birth weight centile and paternal BMI is also presented in Figure 3-4. There was no significant association between paternal BMI and offspring customised birth weight centile (R^2 = 0.002, p = 0.857).

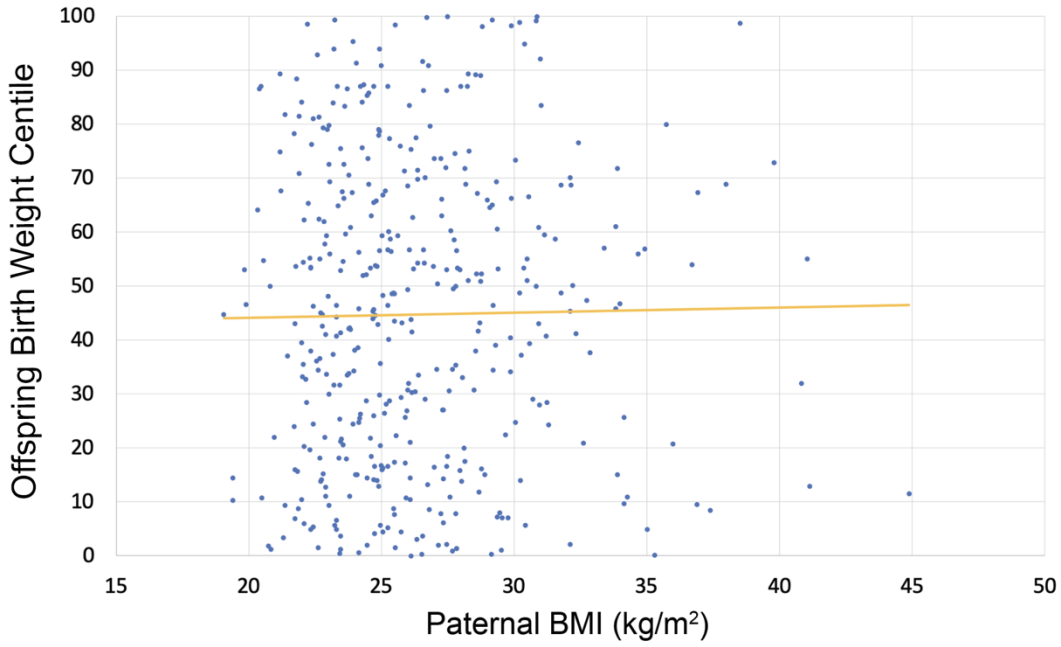


Figure 3-4. Association between paternal BMI and offspring customised birthweight centile

The relationship between paternal insulin resistance and offspring birth weight was further explored by regressing offspring customised birth weight centiles against paternal HOMA-IR (Figure 3-5). There was no significant association between paternal HOMA-IR category and offspring birth weight centile (p = 0.326).

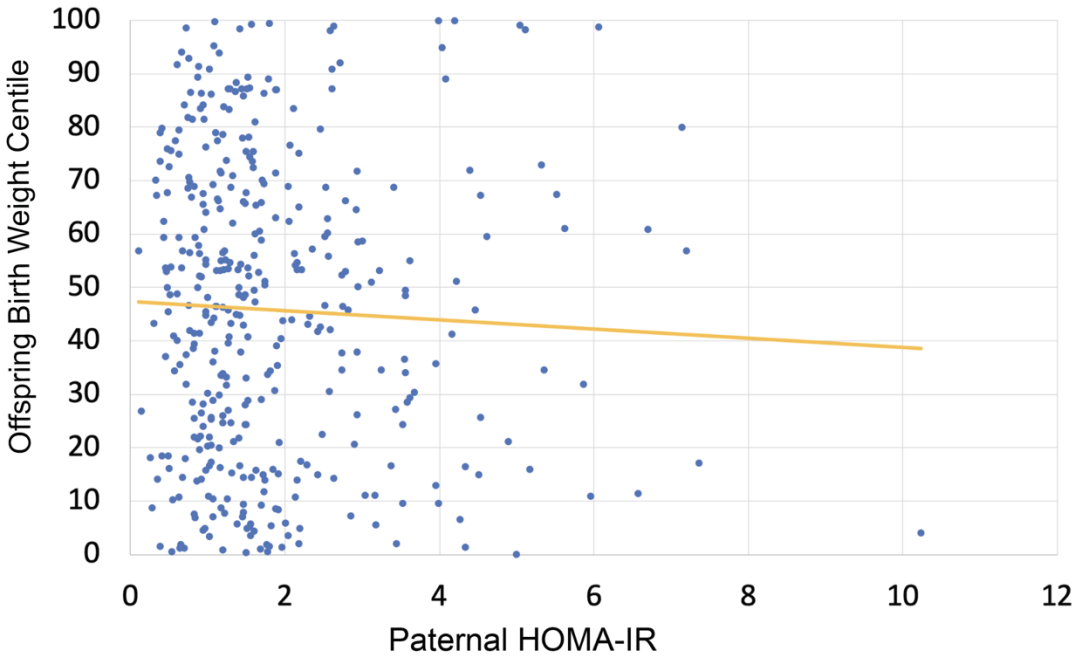


Figure 3-5. Association between paternal HOMA-IR and offspring customised birth weight centile

3.5.6 Remembered Parental Birth Weights

There were 243 out of the 390 (62%) participating fathers and 258 out of the 390 (66%) participating mothers who remembered their own birth weight or acquired this information from a parent. Because few of the parents confidently remembered the length of gestation at which they were born, these parental birth weights were analysed as raw values in grams rather than customised centiles.

There was a trend towards paternal birth weight correlating with offspring birth weight, although this did not reach statistical significance (Table 3-6).

	Parents of SGA	Parents of AGA	Parents of LGA	p
n	48	324	18	
Paternal weight at birth. Mean (SD)	3393 (499)	3548 (592)	3700 (386)	0.231
Fathers who remembered birth weight (%)	69	62	56	
Maternal weight at birth. Mean (SD)	3458 (757)	3268 (572)	3310 (534)	0.240
Mothers who remembered birth weight (%)	69	66	56	

Table 3-6. Remembered parental birth weights according to offspring birth weight centile group

The relationship between parental own weight at birth and offspring weight at birth was further explored by regressing parental birth weight against offspring customised birth weight centiles (Figure 3-6).

Figure 3-6. Association between parental (own) birth weights and offspring customised birth weight centiles
a) Association between paternal (own) remembered birth weight and offspring birth weight centile
b) Association between maternal (own) remembered birth weight and offspring birth weight centile

There was a statistically significant relationship between paternal (own) weight at birth and offspring customised birth weight centile (R^2 = 0.046, p = <0.001). Interestingly, there was no significant association between maternal (own) weight at birth and offspring customised birth weight centile (R^2 = 0.004, p = 0.984) (Figure 3-6).

I thereafter investigated whether fathers that had been small at birth had a higher average BMI and HOMA-IR than fathers who had been born with a normal or high birth weight. Fathers who remembered their birth weights were therefore divided into three groups; those with a birth weight <2.5kg, those with a birth weight 2.5-4.5kg, and those with a birth weight >4.5kg. As presented in Table 3-7, fathers who weighed <2.5kg at birth were heaviest as adults (BMI 27.5 kg/m^2), but this was not significantly different from fathers born >4.5kg (BMI 26.2 kg/m^2). There was no association between paternal own birth weight and adult HOMA-IR.

	Low birth weight fathers	**Normal birth weight fathers**	**High birth weight fathers**	**p (difference between groups)**
n	10	221	16	
Adult BMI. Mean (SD)	27.5 (4.4)	26.1 (4.1)	26.2 (2.4)	0.571
Adult HOMA-IR. Mean (SD)	1.6 (0.7)	1.8 (1.6)	1.8 (1.6)	0.852

Table 3-7. Association between paternal (own) weight at birth and adult BMI and insulin resistance.
Low birth weight fathers refer to men born with a remembered birth weight <2.5kg. Normal birth weight fathers refer to men born with a birth weight of 2.5-4.5kg. High birth weight fathers refer to men born with a remembered birth of >4.5kg.

3.6 Discussion

In this prospective cohort study of 390 mother-father-child trios I identified three groups of fathers according to their BMI (<25, 25-30 and >30 kg/m^2) and followed up their partners' pregnancies. I did not identify any paternal metabolic risk factors for fathering small for gestational age offspring. However, I did identify a significant association between paternal own birth weight and offspring customised birth weight centile. There was no such association between maternal own offspring birth weight centile. I also identified a trend towards higher insulin resistance in fathers of SGA infants, although this association was not significant.

The finding of a significant association between paternal own birth weight and offspring birth weight should be interpreted with caution, as approximately one third of fathers did not remember their birth weight, and this was not a primary outcome of the study. However, the association between paternal own weight at birth and the birth weight of his offspring is in concordance with previous research in this area (67, 68). It is interesting to hypothesise why offspring birthweight was associated with paternal but not maternal (own) birth weight. The observation suggests a paternal genetic influence on offspring birth weight, whilst the maternal influence on offspring's birth weight is predominantly via her intrauterine environment, e.g. nutrient availability. This is in concordance with an extension of the parental conflict theory. The theory posits that paternal (epi)genetic variants are generally growth-promoting, in order to increase the evolutionary fitness of his genes, whereas maternal (epi)genetic variants are generally growth-restricting, as she needs to balance the requirement for a large offspring with conserving resources for her own survival.

Studies that have found an association between paternal metabolic traits and offspring growth may offer clues as to why my results did not support this association. For example, a nested cohort study by McCowan et al involving 2002 couples in the prospective Screening for Pregnancy Endpoints (SCOPE) cohort found that men who fathered SGA infants were 36% more likely to be obese compared to men who fathered normally grown infants (68). However, McCowan et al also found that

fathers of SGA infants were more likely to themselves have been small at birth. In that study, fathers of SGA infants were approximately 180 g lighter at birth compared to men who fathered non-SGA infants ($p = < 0.01$). As previously discussed, there is a well-established relationship between low weight at birth and an increased risk of adult metabolic disease (60). It is also well established that paternal birth weight is a significant and independent predictor of low birth weight in offspring (67). It is possible, therefore, that rather than paternal obesity being causal of low offspring birth weight through the inheritance of epigenetic changes acquired with obesity, paternal genetic variants that predispose to low birth weight may mediate a dual phenotype that also predisposes to metabolic disease in later life.

In line with this, a case-control study by Hillman et al identified that males who had recently fathered growth restricted offspring were more likely to be sub-clinically insulin resistant (OR 7.68 of having a 1 unit higher log HOMA-IR value; 95% CI 2.63–22.40; $p = <0.001$), and to themselves have been lighter at birth (69).

From these two studies would follow that fathers who had themselves been light at birth would be more likely to have a higher BMI and higher HOMA-IR later in life. I did investigate whether there was an association between a low paternal own birth weight and obesity or insulin resistance later in life, however I found no association, potentially because the study was underpowered to detect such effects.

Hillman et al also found that males who had fathered growth-restricted offspring were more likely to smoke cigarettes (OR 3.39; 95% CI 1.26–9.16; $p = 0.016$). A recent meta-analysis of the association between paternal smoking and offspring risk of SGA, comprising a total of 29,366 infants, concluded that paternal smoking was associated with a small but significant increase in the risk of fathering SGA offspring (pooled estimate = 1.21 (95% CI 1.03–1.44) (227). Smoking is itself associated with insulin resistance (234). It is also possible that males who are smokers are also more likely to lead a lifestyle that is otherwise also associated with metabolic disease. Therefore, it is possible that rather than paternal insulin resistance being directly associated with

low offspring birth weight in this cohort, part of the association could be explained by paternal smoking that negatively impacts on fetal growth.

Furthermore, from a statistical perspective, I found that measures of insulin, HOMA-IR and HOMA2-IR in the Dad's health Study were skewed, as opposed to normally distributed. Hillman et al treated the values as conforming to a normal distribution. Thus, the descriptive statistics and statistical analyses employed were slightly different. It is possible that these differences in study group and statistical analysis contributed to different results.

In agreement with findings from the Dad's Health Study, a recent systematic review of the paternal contribution towards perinatal outcomes did not find an association between paternal BMI and offspring birthweight (227). However, this systematic review did find an association between paternal height and offspring birthweight. In the Dad's Health Study, I did not find a significant association between paternal height and offspring birth weight centile ($p = 0.344$), however the study was likely underpowered to detect modest effects.

3.6.1 Strengths and Limitations

The Dad's Health Study is among the first prospective studies of the association between paternal obesity and insulin resistance and offspring birth weight. Previous comparable studies have either been retrospective case-control studies (69), or not included measures of insulin resistance but assessed more crude measures of metabolic health such as BMI alone (68). Therefore, strengths of the Dad's Health Study include that the prospective study design helped to reduce potential selection bias and increase generalisability of results. The detailed phenotypic assessment of male participants also allowed me to obtain a more sensitive measure of, for example, paternal insulin resistance compared some previous studies that have used deaths associated with diabetes as a proxy for metabolic dysregulation (224).

The findings of the Dad's Health study suggest that the influence of paternal obesity and insulin resistance on offspring birth weight is limited. However, a limitation of

the Dad's Health Study is that it was likely underpowered to detect modest effects of paternal metabolic health on fetal growth, in particular across the spectrum of normal birth weights. For example, it is possible that the trend of a higher insulin resistance in fathers of SGA offspring (mean HOMA1-IR = 1.6) compared to fathers of AGA offspring (mean HOMA-IR = 1.4) would have been significant had the sample size been larger. The power calculations performed prior to commencing study recruitment were based on recruiting lean and obese fathers only. However, the actual study also included a group of overweight fathers. This could have reduced my ability to identify significant effects. Further, the power calculations were based on previous retrospective studies in this field which might have included more severe phenotypes and therefore required smaller numbers of participants to demonstrate an effect (69). In particular, the Dad's Health Study included a small number of SGA offspring (n = 48). Although this number is comparable to that in the study by Hillman et al (n = 42), it is likely that the SGA trios I studied included a heterogenous group of both growth-restricted and constitutionally small neonates. For comparison, the mean customised birthweight centile in the study by Hillman et al was 1.8 (SD = 2.2) and the mean customised birth weight centile among SGA infants in the Dad's Health Study was 4.6 (SD = 3.1). Constitutionally small neonates are less likely to be affected by metabolic dysregulation than are those affected by growth restriction.

It is also possible that despite controlling for maternal variables, e.g. by using customised birth weight centiles and excluding mothers who developed gestational diabetes, that could still be residual maternal confounding influencing the results.

Another limitation of the study comes from the nature of recruitment to the Dad's Health Study. The majority of male participants were recruited after responding to a study poster or a study leaflet placed in antenatal waiting areas at UCLH. The incentive for study participation was that metabolic health data such as blood pressure, insulin and glucose and blood lipid measurements would be shared with participants following their study visit. It is possible that men who were more interested in receiving such information were more likely to participate. This would reduce the generalisability of my findings.

A potential source of error in the Dad's Health Study imprecision in how infants were classified as SGA, AGA and LGA respectively. There is considerable controversy in how such a classification should occur (228, 235, 236). For example, some large-scale population based studies of 'optimal' fetal growth (that is, fetal growth not obviously compromised by maternal factors such as illness or poor nutrition) across several geographical regions have led to the development of birth weight standards proposed to be used across various geographical settings and without taking individual maternal factors, such as maternal BMI, into account (235). Other have argued for the use of customised birth weight centiles, which control for factors such as maternal BMI, ethnicity and parity, as these may more accurately reflect the individual fetus' ability to meet his growth potential (228). It has been argued that the widespread adoption of customised birth weight centiles across the UK has improved the detection of at-risk fetuses and reduced the rates of stillbirths due to this ability to individually assess the ability of each fetus to meet their growth potential (228).

The Dad's Health Study used customised birth weight centiles rather than birth weight standards that do not take individual maternal characteristics into account, such as Intergrowth-21, for several reasons (235). First, comparable retrospective studies had employed customised birth weight centiles, and using the same methods allowed study results to be compared more readily (69). Further, as customised birth weight centiles are now extensively employed across the majority of UK hospitals, it was felt that results from the Dad's Health Study would be easier to interpret in a wider context of obstetric care by using the same definitions as used in clinical practice (228).

One approach to reduce the influence of assortative mating and maternal BMI on study findings would have been to exclude all mothers with a BMI exceeding 25 kg/m^2. However, the feasibility of recruiting adequate numbers of study participants during the time period given would have been considerably compromised by such an approach. For example, in the present study, had we excluded all mothers with a pre-pregnancy BMI of 25 kg/m^2, we would have approximately halved the number of

obese male participants in the study (29 out of 64 would have been excluded). We therefore opted to instead control for the influence of maternal BMI via the use of customised birth weight centiles. As an aside, although numbers are too limited to draw definitive conclusions, I did assess whether there was an association between paternal BMI and offspring customised birth weight centiles when including lean mothers only (n = 278). However, regressing offspring customised birth weight centile onto paternal BMI did not yield a statistically significant finding (R^2 = 0.005, p = 0.115).

3.6.2 Future Directions

As shown in a recent systematic review investigating the paternal role in offspring development, the association between paternal BMI and offspring health may not be evident when studying offspring birth weight alone, but rather when studying offspring weight in later childhood (227). Future studies should explore this possibility by longer term follow up of infant growth and metabolism born to parents with different metabolic phenotypes. This type of research is difficult in humans due to the strong effect of the shared, possibly 'obesogenic', environment of children and parents. It could nevertheless give clinically important insights into how targeted public health strategies could be directed towards families with an increased risk of developing obesity and T2DM.

It would also be interesting to explore whether growth-restricted infants born to parents with metabolic disease are particularly vulnerable to metabolic consequences of rapid weight gain in their early years, comparable to the health consequences of 'catch-up growth' that is associated with metabolic dysregulation in infants born with low birth weight (66).

It is possible that paternal metabolic dysregulation may affect fetal growth in a sex-specific manner. For example, an observational study indicated that paternal BMI is associated with offspring birth weight in male but not in female infants (237). With only 23 female and 25 male infants born SGA, the Dad's Health Study was not powered to detect sex-specific differences in the paternal impact on fetal growth.

Instead, this study used customised birth weight centiles that adjusted for the impact of sex on birth weight. However, future research should further investigate the potential sex-specific impact of the paternal metabolic health on offspring birth weight.

3.7 Summary

This prospective cohort study of 390 mother-father-offspring trios did not identify any paternal metabolic traits that were significantly associated with an increased risk of fathering SGA offspring. It is possible that despite adjusting for maternal BMI, assortative mating underpowered my ability to identify an association between paternal insulin resistance and fathering a growth restricted offspring. Longer term follow-up of the neonates may also reveal paternal influences on childhood growth and metabolism. However, in line with previous research, I identified a significant association between paternal (own) birth weight and offspring birth weight centile. This observation supports the notion that paternal genotype is more influential in determining his offspring birth weight compared with acquired traits such as obesity.

Chapter 4
DNA Methylation in Human Sperm: A Systematic Review

4.1 Introduction

Adequate epigenetic regulation of germ cells is imperative for embryogenesis and offspring health. Epidemiological studies in humans suggest that environmental and acquired paternal traits such as advanced age and smoking have the potential to negatively impact on the development and physiology of his offspring, presumably via alterations to his spermatozoa (227). The sperm methylome in non-human mammals can be altered by environmental and physiological changes, including dietary alterations, toxins, and even psychological stress (132, 134-137, 238). In addition, acquired traits in male mice induce epigenetic changes in his sperm which influence the physiology of his offspring (127, 140). It is possible that the same is true for men.

There is limited, if any, evidence for such germline epigenetic inheritance in humans. Some researchers have suggested that similarities between epigenetic profiles of paternal and offspring blood can be seen as evidence for paternal germline transmission of epigenetic markers (153). However, such studies fail to account for the widespread epigenetic resetting of the genome that occurs shortly after fertilisation and during gonadal development in the human embryo (128).

Human sperm is a less readily accessible tissue than peripheral blood and is only routinely analysed within reproductive medicine settings. This explains why our understanding of the human sperm epigenome is significantly less detailed compared with many somatic tissues. However, in the context of understanding how paternal environmental and acquired traits could influence offspring phenotypic outcomes, sperm represents the most relevant tissue to study. The growing interest in the field of inter- and transgenerational epigenetic inheritance partly underlies the increasing number of studies investigating the DNA methylation landscape of human spermatozoa. Another important reason for such research is the high prevalence of sub- and infertility. These conditions affect approximately one in seven UK couples and is attributed to male factor infertility in approximately 50% of cases (239, 240). Male factor infertility is associated with abnormal semen parameters, such as low sperm count, on routine semen analysis. However, the predictive power of semen

analysis tests is low (241). For example, a study that assessed semen parameters (sperm concentration, motility and morphology) in males in 765 subfertile couples and 696 fertile couples found that there was extensive overlap in sperm parameter measurements between the two groups (241). In other words, while a low sperm concentration was more frequently seen in the subfertile groups, none of the measurements was a powerful discriminator for fertility (241). Also, genetic variants linked to subfertility are only able to explain a small proportion of cases (242). Male factor infertility has, however, been epidemiologically linked to a range of lifestyle and environmental factors, including cigarette smoking, obesity and toxins such as organophosphates (239, 243). These associations make a plausible case for male infertility being at least in part caused by gene-environment, or epigenetic changes.

With this background, the last couple of decades have seen a number of studies investigating the DNA methylation landscape of human spermatozoa. However, findings have rarely been replicated across studies. This is likely due to limited study sizes and results generated using different methodologies. Thus, it has been challenging to draw overall conclusions regarding the human sperm methylome and its potential to change in response to environmental or physiological alterations.

In order to summarise current knowledge, generate recommendations for future research as well as inform the next stages of my project, I undertook a systematic review, following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (191).

4.2 Declarations

The search for relevant papers, filtering of results and summarising of studies was first performed by myself and then repeated independently by Dr Amy Danson, Queen Mary University London (QMUL). Any differences in the selection of articles between myself and Dr Amy Danson were resolved by discussion. Dr Amy Danson also added to the grading of studies and together we discussed overall conclusions and recommendations for future research. Dr Sarah Marzi, QMUL, had an advisory role in the search process and evaluation of evidence.

4.3 Specific Objectives

1. To conduct a comprehensive literature review of studies that investigated DNA methylation of human spermatozoa that were published between 01/01/2003 and 31/03/2019, in accordance with PRISMA guidelines.
2. To summarise studies included in the systematic review with regards to methodologies, limitations and main findings.
3. To objectively evaluate evidence from studies included in the systematic review using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system (244).
4. To generate recommendations for future research.

4.4 Methods

The search strategy, selection of articles and criteria for evaluation evidence are discussed in detail in Chapter 2 section 2.9.

Briefly, publications relating to studies of DNA methylation in human sperm published between 01/01/2003 and 31/03/2019 were identified from the PubMed, Web of Science and Cochrane databases using the following search criteria:

"semen" [Mesh] OR "sperm" [Mesh] AND "DNA methylation" [Mesh] OR "methylome" [Mesh]

Articles identified through the search, but which did not meet the inclusion criteria, were recorded along with their reason for being excluded.

The quality of evidence in articles selected for the systematic review was objectively rated according to the GRADE criteria as described in detail in Chapter 2 section 2.9.3. A summary of the criteria is also copied below for clarity (Table 4-1).

GRADE criteria	Rating	Reasons for down- or upgrading	Quality of the evidence
Study design	RCT (High) Non-RCT (Low)	Only non-RCTs	High Moderate Low Very low
Risk of Bias	No Serious (-1) Very serious (-2)	Age Smoking Somatic cell contamination Storage time/sample storage conditions Medication/supplement use Drug and alcohol use Abstinence Medical history of participants BMI/ diet	
Inconsistency	No Serious (-1) Very serious (-2)	Lack of replication cohort Similar studies showing contradicting results	
Indirectness	No Serious (-1) Very serious (-2)	Generalisability Cross-sectional/longitudinal	
Imprecision	No Serious (-1) Very serious (-2)	Small sample sizes Lack of appropriate controls Varying sperm processing Definition of subfertility Lack of descriptive statistics of participants	
Publication Bias	Undetected Strongly suspected (-1)	Candidate gene approaches	
Other (upgrading factors)	Large effect (+1 or +2) Dose response (+1 or +2) No plausible confounding (+1 or +2)		

Table 4-1. GRADE criteria as adapted for the systematic review of studies of DNA methylation in human sperm
RCT = Randomised Controlled Trial. Adapted from (244).

4.5 Definitions of Sperm Analysis Parameters

The most common reference guide for semen analysis is the 'WHO laboratory manual for the examination and processing of human semen' (245). Unless otherwise stated, 'abnormal semen parameters' refers to samples that have parameters below the 5th centile as defined by this manual (see below). The current version of the manual was published in 2010. Therefore, some of the earlier studies discussed in this review employed an earlier version of the manual, which was one published in 1999. However, I have not regarded the differences in the definitions adopted by these two versions of the manual as significant enough to prohibit comparisons between studies.

The semen parameters most commonly analysed during a routine semen analysis include sample volume, appearance, pH, viscosity, sperm motility, viability/vitality (i.e. the percentage of live sperm, which is particularly relevant if motility is low), concentration and morphology. The total sperm count is regarded as the most accurate indication of whether a semen sample is normal, and is calculated by multiplying the sample volume and concentration (245). Other tests, such as the mixed antiglobulin reaction (MAR) test, is included in a semen analysis when indicated. The MAR test assesses the presence of antibodies on sperm heads. A high level of these anti-sperm antibodies is associated with decreased fertility (245).

Some specific definitions that appear in the systematic review are presented below. All definitions are derived from the 'WHO laboratory manual for the examination and processing of human semen' (245).

Asthenozoospermia: reduced sperm motility. This is determined by the percentage of progressively motile spermatozoa, i.e. spermatozoa which move at a speed of 25 µm per second or more. The 5th centile for progressive motility is 32% progressively motile spermatozoa. Below this level, the sample is considered asthenozoospermic.

Azoospermia: no sperm in the ejaculate.

Normozoospermia: a normal semen sample. Total number of spermatozoa, and percentages of progressively motile and morphologically normal spermatozoa, equal to or above the lower reference limits. This means:

- Volume ≥ 1.5 ml
- Concentration ≥ 15 million sperm/ml
- Total number of sperm ≥ 39 million
- Progressive motility ≥ 32%
- Vitality ≥ 58%
- Morphology ≥ 4% of sperm with normal morphology

Oligozoospermia: a low number of spermatozoa in the ejaculate. Defined as less than 39 million sperm/ejaculate

Teratozoospermia: sperm with abnormal morphology. A teratozoospermic sample has less than 4% morphologically normal sperm

4.6 Results

4.6.1 Articles selected for inclusion

The selection process of articles for inclusion in the systematic review is presented in Figure 4-1.

A total of 463 articles were identified by the search terms and through manual filtering of reference lists. Two of these were duplicate records and were thus removed. We excluded 120 records as they were reviews or did not describe original research. A further 217 articles were excluded because the study did not investigate humans (n = 51), investigated another tissue than sperm (n = 86), related to forensic identification of semen (n = 29) or was published before 2003 (n = 15). This resulted in 124 studies that met the predetermined inclusion criteria and were included in the systematic review. To my knowledge, these articles represent all published studies investigating DNA methylation in human sperm that were available up to March 2019.

Of the 124 articles included in the systematic review, 17 investigated global DNA methylation, 61 investigated candidate genes, and 46 were genome-wide analyses. These three types of investigations were evaluated separately and are presented below.

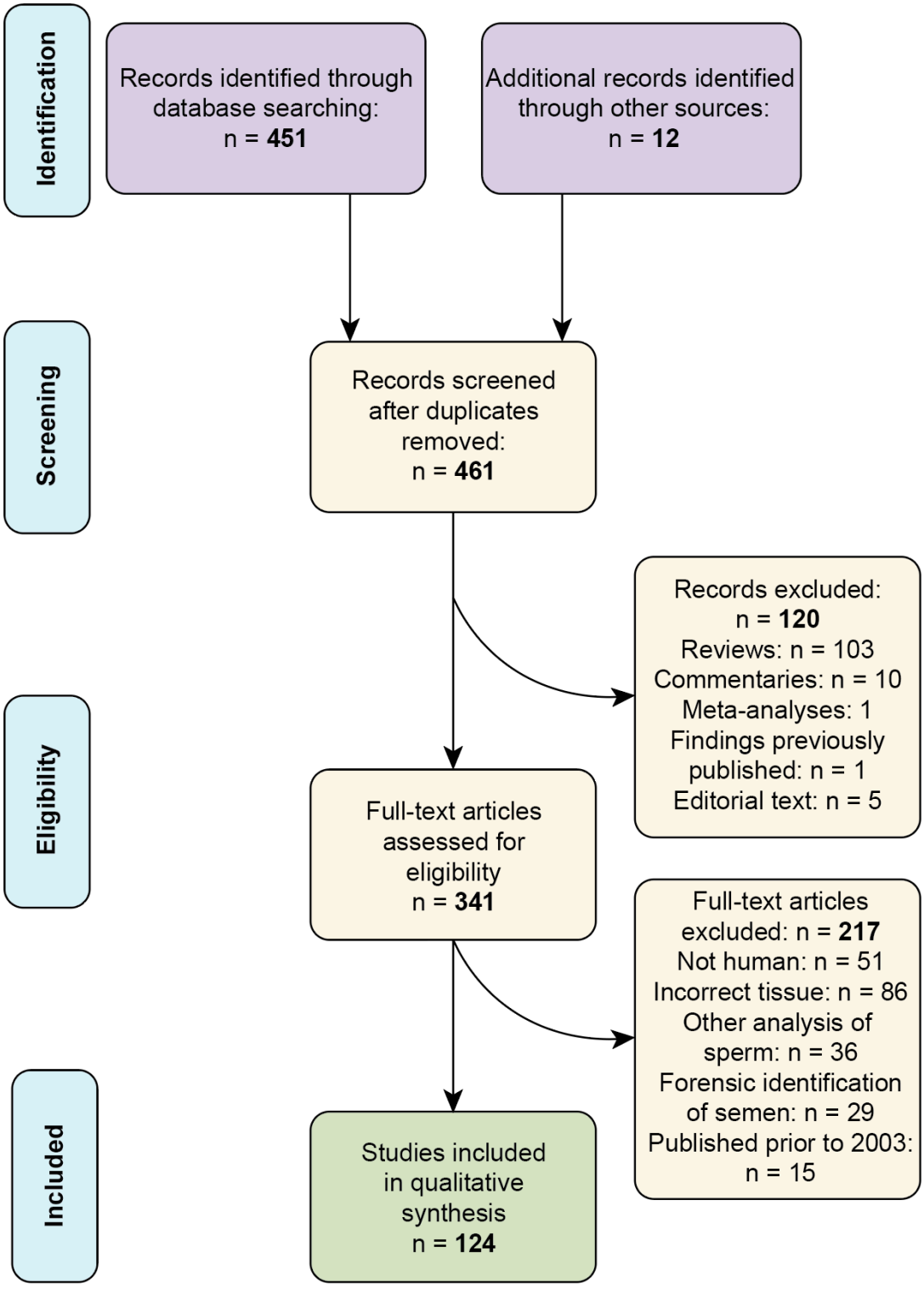


Figure 4-1. Flow chart of selection of records for the systematic review
Assembled according to PRISMA guidelines (191).

4.7 Studies of Global DNA Methylation in Human Sperm

Studies of global sperm DNA methylation are presented in Table 4-2.

Global DNA methylation refers to the total level of methylation across the entire genome. The gold standard for quantifying global levels of 5mC is liquid chromatography coupled with mass spectrometry (246). However, the cost and requirement for a mass spectrometer is often prohibitive. Therefore, immunohistochemical staining with antibodies against 5mC followed by a secondary antibody coupled to a reporter molecule such as fluorescein isothiocyanate for microscopic quantification of 5mC is a commonly used method (247, 248).

Analyses of global DNA methylation have historically been most informative in oncology, where global hypomethylation has been found to be characteristic of several forms of cancer (246). However, the significance of global alterations of DNA methylation in spermatozoa is yet unclear. Indeed, several of the studies included in this review failed to provide a biological hypothesis justifying the analysis of global DNA methylation in association with particular phenotypes.

Sample sizes varied widely but were often limited; the median number of total participants (cases and controls combined) was 60 (range 10-316) and the mean 99.

Methodologically, most studies (10/17) employed some form of immunostaining for 5mC as an assay for detecting global DNA methylation. However, this semi-quantitative technique relies on visual image analysis and is thus prone to e.g. observer variability. Also, the technique is prone to underestimating levels of 5mC as the 5mC modified bases can be hidden within the double-stranded DNA helix (248). In 7 out of 17 studies, the method of cleaning the sample from somatic cells was either not detailed, or only involved washing the sample in phosphate buffered saline (PBS). Thus, results from these studies may have been influenced by somatic cell contamination.

The quality of evidence for the majority of studies was graded as low or very low (11 out of 17). The main reasons for downgrading the quality of evidence included a lack of correction for covariates, lack of correction for multiple testing, use of immunostaining as the sole method of assaying global DNA methylation, and lack of appropriate removal of somatic cells. In addition, the generalisability of findings to the general male population was found to be overall low, with approximately one third of studies (6/17) recruiting participants solely from reproductive medicine settings. The studies where the quality of evidence was graded as high generally included taking account of important epigenetic confounders such as age and smoking status, and validating global 5mC measurements by using more than one technique, e.g. as done by Barzideh *et al.* (249).

Author, year, reference	*Type of study*	*Area of research*	*Method of assaying DNA methylation*	*Initial processing*	*N (cases)*	*Age (years); description of cases*	*N (controls)*	*Age and description of controls*	*Main Results*	*Study Quality*
Benchaib et al., 2003 (247)	Cohort	Fertility/ ART	Immuno-staining for 5mC	PureSperm gradient (50, 70 and 90%)	23	35 (median); males undergoing ART	None	N/A	Global 5mC lower in teratozoospermia than non-teratozoospermia	Very Low
Benchaib et al., 2005 (250)	Cohort	Fertility/ ART	Immuno-staining for 5mC	PureSperm gradient (50, 70 and 90%)	63	35.4 (+/- 5.5); males under-going ART	None	N/A	Increased global 5mC associated with higher rates of pregnancy and embryo development, but not with fertilisation rates	Low
Aoki et al., 2006 (251)	Cross-sectional	Fertility/ ART	Quantitative immune-fluorescence microscopy for 5mC	ND	195	Age ND; sub-fertile males (divided into 3 groups based on P1/P2 ratios)	None	N/A	No significant differ-ences in global DNAm between groups of varying protamine P1/P2 ratios	Low
Tavalaee et al., 2009 (252)	Cohort	Fertility/ ART	Immuno-staining for 5mC	PureSperm gradient (40 and 80%)	61	36.2 (+/- 5.2); males under-going ART	None	N/A	Global DNAm negatively correlated with DNA fragmentation	Moderate
Tunc, Tremellen, 2009 (253)	Cohort	Medic-ations	Immuno-staining for 5mC	Ultra-centrifugation at 300g for 20 minutes	45	Age ND; sub-fertile males (two samples each)	12	Age ND; proven fertility	3 months of antioxidant supplementation asso-ciated with a fall in seminal ROS levels and sperm DNA fragment-ation, and increased sperm DNAm	High

Barzideh et al., 2012 *(249)*	Cross-sectional	Fertility/ ART	High-pressure liquid chroma-tography for 5mC. Valid-ated with flow cytometry	Percoll gradient	16	20-23; normo-zoospermic	N/A	N/A	5mC lower in high-density portion of spermatozoa than in low-density portion	High
Jenkins et al., 2013 *(254)*	Cohort	Aging	Immuno-staining for 5mC; 450K array of paired samples from 2 'repre-sentative' donors	Somatic cell lysis	67	15 for DNAm and ageing study (2-3 samples each, years apart); 52 for blood-sperm comparison	N/A	N/A	Significant increase of global 5mC and 5hmC levels with age. 5hmC levels in sperm generally lower than in blood	Low
Leter et al., 2014 *(255)*	Cross-sectional	Toxins	Targeted bisulfite-specific PCR and flow cytometric immuno-detection of 5mC	ND	262	29.2 (+/- 0.3); 71 from Green-land, 96 from Poland and 95 from Ukraine	N/A	N/A	No consistent associations between internal PFASs concen-trations and any of the methylation biomarkers	Moderate
Consales et al., 2014 *(256)*	Cross-sectional	Diet/ lifestyle/ meta-bolic disease	Targeted bisulfite-specific PCR and flow cytometric immuno-detection of 5mC	ND	316	29.3 (+/- 0.3); 116 from Greenland, 100 from Poland and 100 from Ukraine	N/A	N/A	Geographical location (Greenland versus European towns) identified as the main determinant DNAm in repetitive sequences	Moderate

Bahreinian et al., 2015 *(257)*	Cross-sectional	Fertility/ ART	Flow cytometric immuno-detection for 5mC	Samples washed in PBS	44	31.1 (+/- 0.6); Males with grades II and III varicocele	15	37.8 (=/- 2.0); proven fertility	Varicocele associated with lower global 5mC and lower sperm parameters	Very low
Yu et al., 2015 *(258)*	Cross-sectional	Method-ology	MeDIP-chip assay and immuno-staining for 5mC	v/v medium gradient (40 and 80%)	54	32.4 (+/-5.2); Astheno-zoospermic or oligoasthenozoo spermic	39	33.6 (+/- 4.8); normo-zoospermic	Global 5mC levels lower in motile sperm selected by density centrifugation compared to the less motile sperm	Low
Kim et al., 2015 *(259)*	Cross-sectional	Toxins	Immuno-staining for 5mC	Swim-up	18	32.0 (+/- 5.5); Smokers	32	Age ND; non-smokers	5mC levels lower in swim-up fraction of spermatozoa	Low
Tavalaee et al., 2015 *(260)*	Cohort	Fertility/ ART	Immuno-staining for 5mC	Samples washed in PBS	23	31.3 (+/- 4.3); Males with grade II and III varicocelectomy (one sample before and one after surgery)	None	N/A	Varicocele surgery associated with improved sperm parameters, but no significant change in global DNA methylation	Low
Cassuto et al., 2016 *(261)*	Cross-sectional	Methodo logy	Immuno-staining for 5mC	Sperm separation medium gradient (45 and 90%)	10	37-50; 448 S6 spermatozoa compared to 428 S0 spermatozoa	N/A	N/A	Lower global 5mC in morphologically higher scoring spermatozoa compared to morpho-logically lower scoring spermatozoa	Low

Consales et al., 2016 *(262)*	Cross-sectional	Toxins	Flow cytometric immuno-detection of 5mC. DNAm levels at *Alu, LINE-1, Satα* assayed by PCR pyro-sequencing	Samples washed in PBS	269	29.22 (+/- 3.2); 75 samples from Greenland, 97 from Poland and 97 from Ukraine	N/A	N/A	Flow cytometry suggested association between persistent organic pollutant (PCBs or DDT) exposure and lower global DNAm	High-moderate
Uppangala et al., 2016 *(263)*	Cohort	Fertility/ ART	Immuno-staining for 5mC	Swim-up	19	Age ND; proven fertility. Samples provided after 1, 3, 5 and 7 days of abstinence	N/A	N/A	No significant association between 5mC levels and length of ejaculatory abstinence	Very low
Olszewska et al., 2017 *(264)*	Cross-sectional	Fertility/ ART	Thin-layer chromato-graphy and immuno-fluorescence techniques to assess global levels of 5mC	Samples washed in F10 medium	9	30.7; subfertile carriers of chromosomal abnormalities	14	28.5; healthy volunteers	Chromosomal abnormalities not associated with overall difference in 5mC levels	Very low

Table 4-2. Studies of global DNA methylation in human sperm.
'Gradient' refers to a discontinuous gradient ultracentrifugation using the medium detailed in the table. 'Initial processing' refers to how the semen sample has been cleaned from seminal fluid, debris, somatic cells and immotile/dead spermatozoa. ND = Not Detailed, 5mC = 5-methylcytosine.

4.7.1 Summary of findings from high quality studies of global DNA methylation

Three studies of global sperm DNA methylation were assessed as being of high quality. One of these suggested that antioxidant supplementation could be linked to a fall in levels of reactive oxygen species (ROS) in semen and an a concomitant increase in global sperm DNA methylation (253). The authors suggest that ROS increases sperm DNA fragmentation, which leads to overall hypomethylation (253). Normal spermatozoal gene expression is characterised by overall transcriptional repression (265). In general, transcriptional repression is linked to high levels of methylation, especially of promoter regions (266). It is possible that ROS or other toxins damage DNA integrity with a resulting decrease in global DNA methylation. In line with this, another high quality study of global DNA methylation suggested that exposure to persistent organic pollutants could be associated with global hypomethylation (262). In contrast to this, however, a third high quality study by Barzideh et al found that DNA methylation levels were generally higher in the low quality proportion of spermatozoa (249). The authors suggested that an abnormally hypermethylated state was associated with the pre-apoptotic state (by using annexin V binding as a proxy for tendency towards apoptosis) (249). These results are not necessarily contradictory; it is possible that whilst normal spermatozoa display a relatively high level of DNA methylation that can be lowered by toxins such as ROS and persistent organic pollutants, this level of DNA methylation could be increased above the normal during the process of apoptosis.

4.8 Candidate Gene Analyses of DNA Methylation in Human Sperm

Candidate gene analyses of human sperm methylation are presented in Table 4-3.

Early studies of the human sperm methylome followed epidemiological reports of a possible association between assisted reproductive technologies and an increased risk of imprinting disorders in the offspring (154). It was hypothesised that spermatozoa from males with abnormal semen parameters who underwent in vitro fertilisation or intracytoplasmic sperm injection could harbour a greater frequency of abnormally imprinted genes that would impact on the phenotype of the offspring. Genomic imprints are established on gametes, persist during embryonic

development and adulthood, and erased and re-established during gonadal development in germ cells (in the next generation) (267). It is therefore possible, at least in theory, that if these imprinted genes are inaccurately established on gametes, they could influence offspring development.

With this background, several early candidate gene analyses were directed towards the imprinted gene cluster at 15q11-q13, which is implicated in the Prader-Willi and Angelman syndromes, as well as the *H19/IGF2* locus, which is implicated in the Beckwith-Wiedemann and Silver-Russell syndromes (159, 268-271). Since then, candidate gene analyses have continued to, in particular, analyse DNA methylation profiles of imprinted genes, as well as genes implicated in processes such as germ cell function and embryonic development.

4.8.1 General features of candidate gene analyses of human sperm

For the majority of studies (32/61), the grade of evidence was deemed to be low or very low. The main reasons for downgrading studies was lack of adjusting for covariates, lack of description of phenotypic traits of participants (e.g. making it unclear whether cases and controls were appropriately matched), and lack of correction for multiple testing. Sample sizes varied widely; the median total number of participants (cases and controls combined) was 58 (range 10-350) and the mean was 85.

Further, there were considerable methodological differences in the processing of semen samples prior to methylation analyses, making cross-study comparisons difficult. Thus, around one fifth of studies (12/61) used a swim-up method for selecting motile sperm, around one sixth (9/61) used a discontinuous gradient ultracentrifugation protocol and three studies used a combination of swim-up and a discontinuous gradient ultracentrifugation protocol. In around one quarter of the studies (16/61), no sperm selection method was detailed. In a further five studies the semen samples were centrifuged or washed in PBS, but no sperm selection method was applied. For such studies it is difficult to see that appropriate care had been taken not to analyse DNA methylation of somatic cells such as leukocytes and epithelial

cells. The remaining three studies used a somatic cell lysis buffer to remove somatic cells, or handpicked spermatozoa. In general, there was a lack of validation that the sperm selection methods had been successful (e.g. by microscopic examination of cleaned samples or bioinformatically adjusting for cell type composition).

Author, year, reference	*Study design*	*Area of research*	*Method of assaying DNA methylation*	*Initial processing*	*N (cases)*	*Age (years); description of cases*	*N (controls)*	*Age (years); description of controls*	*Main Results*	*Study Quality*
Marques et al., 2004 *(272)*	Case-control	Fertility/ ART	Bisulfite sequencing	Gradient centrifugation and swim-up	96	Age not detailed; oligozoospermic	27	Age not detailed; normozoospermic	*H19* abnormally imprinted in oligozoospermic group	Moderate
Grunau et al., 2005 *(273)*	Cross-sectional	Tissue specificity	Methylation specific PCR	Not detailed	4	Not detailed	N/A	N/A	Relative hypo-methylation of *BAGE* in spermatozoa and malignant tissues compared to normal, somatic tissues	Very low
Stöger et al., 2006 *(274)*	Cross-sectional	Diet/ lifestyle/ metabolic disease	Hairpin-bisulfite PCR	Not detailed	1	Not detailed	N/A	N/A	The *LEP* CpG island generally unmethy-lated in both human and mouse sperm compared to somatic tissues	Very low
Li et al., 2006 *(275)*	Cross-sectional	Fertility/ ART	Bisulfite sequencing	Samples washed in PBS	4	Age not detailed; proven fertility	N/A	N/A	*DAZ1* promoter methylated in leuko-cytes but unmethy-lated in sperm	Very low
Kobayashi et al., 2007 *(169)*	Cohort	Fertility/ ART	Combined bisulfite PCR restriction analysis and bisulfite sequencing	Swim-up method	18	27-50; oligozoospermic	79	Age not detailed; normozoospermic	DNAm errors more prevalent in oligo-zoospermia. No significant association between imprinting errors and ART outcome	Very low

Geuns et al., 2007 *(276)*	Cross-sectional	Fertility/ ART	Methylation specific PCR	Percoll gradient (45 and 90%)	Not detail-ed	Not detailed	None	N/A	Intergenic DMR of *DLK1-GTL2* relatively hypermethylated compared to somatic tissues (e.g. oocytes and preimplantation embryos)	Very low
Marques et al., 2008 *(277)*	Case-control	Fertility/ ART	Bisulfite sequencing (Sanger)	Suprasperm gradient (55, 80 and 90%) and swim-up	20	Age not detailed; 6 oligozoo-spermic; 14 asthenozoo-spermic; all abnormal morphology	Not detailed	Age not detailed; normozoosp ermic	*H19* and *MEST* abnormally imprinted in severely oligozoo-spermic group	Moderate
Poplinski et al., 2009 *(278)*	Case-control	Fertility/ ART	Targeted bisulfite Sanger sequencing	Swim-up	148	35.5 (32-41); subfertile males	33	37.0 (31.5-40); normo-zoospermic; infertility attributed to female side	*MEST* hypomethylation significantly associated with oligozoospermia in subfertile males	High
Kobayashi et al., 2009 *(279)*	Cohort	Fertility/ ART	Combined bisulfite PCR restriction analysis and bisulfite sequencing	Not detailed	78	Age not detailed; samples paired with aborted conceptuses conceived via ART	38	Age not detailed; samples paired with aborted conceptuses not conceived via ART	Imprinting errors more prevalent in oligozoo-spermic samples. Imprinting errors on sperm also evident in the resulting concep-tuses in 7 out of 17 cases	Very low

Marques et al., 2010 *(280)*	Case-control	Fertility/ ART	Bisulfite sequencing	Testicular biopsies washed in sperm prepar-ation medium and somatic cells removed with erythro-cyte lysing buffer	24	22-44; azoospermic (sperm obtained from testicular biopsies)	Not detailed	Age not detailed; normo-zoospermic	*H19* DNAm errors identified in azoospermic samples obtained via testicular biopsy	Moderate
Wu et al., 2010 *(281)*	Case-control	Fertility/ ART	Targeted bisulfite Sanger sequencing	Swim-up	94	29.04 (+/- 4.52); idiopathic subfertility	54	29.52 (+/- 3.72); proven fertility	Higher DNAm of the *MTHFR* promoter in subfertile males compared with fertile controls	High
Boissonnas et al., 2010	Case-control	Fertility/ ART	Targeted pyro-sequencing	Percoll gradient (45, 60 and 90%)	41	36.6 (+/- 5.7); divided into teratozoo-spermic, oligo-zoospermic and asthenozoo-spermic groups	17	Age not detailed; normo-zoospermic	Relative hypo-methylation of various CpG positions in the *H19-IGF2* DMR	Moderate
Navarro-Costa et al., 2010	Case-control	Fertility/ ART	Bisulfite sequencing	PureSperm gradient (40 and 80%)	5	39.4 (+/-7.2); oligozoospermic	5	39.2 (+/- 7.3); normo-zoospermic	Hypermethylation of the *DAZL* (but not *DAZ*) promoter more prevalent in oligozoo-spermic samples compared to normo-zoospermic controls	Moderate

Hammoud et al., 2010 *(282)*	Case-control	Fertility/ ART	Targeted Sanger sequencing	Not detailed	20	Age not detailed; 10 with abnormal protamine 1 to protamine 2 ratio and 10 with oligozoospermia (<10 x10^6/ml) (9 and 8 respectively for LIT1)	Not detailed	Age not detailed; proven fertility	Relative hypermethylation of *LIT1, SNRPN, MEST, ZAC, PEG3* in oligozoospermia and abnormal P1/P2 ratio samples compared to normozoospermic samples	Moderate
Nanassy, Carrell, 2011 *(283)*	Case-control	Fertility/ ART	Bisulfite pyro-sequencing	Not detailed	92	28-33; 60 with abnormal protamine P1/P2 ratio, 32 with oligozoospermia	40	32-33; normo-zoospermic, proven fertility, normal protamine ratio	Significantly higher rate of *CREM* promoter methylation in males with abnormal prota-mine P1/P2 ratio and in oligozoospermia compared to fertile, normozoospermic controls	Moderate
Minor et al., 2011 *(284)*	Case-control	Fertility/ ART	Bisulfite sequencing	Swim-up	18	37.8 (+/-6.9); azoospermic (sperm obtained from testicular biopsies)	9	34.1 (+/-2.4); proven fertility (from ejaculate)	Significant hypo-methylation of *H19* in azoospermic males	Very low
Chorbov et al., 2011 *(285)*	Case-control	Toxins	Targeted bisulfite pyro-sequencing	Not detailed	13	42.5; males with opioid addiction	21	29.3; healthy volunteers	Significant correlation between *OPRM1* methylation and opioid dependence in blood but not in sperm	Low

Nanassy, Carrell, 2011 *(286)*	Case-control	Fertility/ ART	Bisulfite sequencing	Not detailed	10	Age not detailed; abnormal protamine P1/P2 ratio	10	Age not detailed; normo-zoospermic	No significant DNAm differences in the of selected genes between samples with an abnormal protamine P1/P2 ratio and normozoospermic controls	Very low
Ibala-Romdhane et al., 2011 *(287)*	Cohort	Fertility/ ART	Targeted bisulfite Sanger sequencing	Gradient ultra-centrifugation (medium not specified)	11	Age not detailed; abnormal semen parameters (mix of oligo-, astheno- and terato-zoospermia	Not detailed	Age not detailed; normo-zoospermic, proven fertility	Embryos with developmental failure more likely to have abnormal imprinting at *H19*, however all corresponding sperm in study had normal imprinting at *H19*	Very low
Netters-heim et al., 2011 *(288)*	Cross-sectional	Cancer	Bisulfite Sanger sequencing	PureSperm gradient (40 and 80%)	Not detai-led	Not detailed	N/A	N/A	*NANOG* promoter hypomethylated in spermatogonia and in several germ cell tumours, but hypermethylated in adult sperm	Very low
Sato et al., 2011 *(289)*	Cross-sectional	Fertility/ ART	COBRA and methylation-specific PCR with Luminex technology	Swim-up	128	Age not detailed; oligozoospermic	209	Age not detailed; normo-zoospermic	High correlation between DNAm analysis methods. Higher levels of imprinting errors in oligozoospermia	Low
El Hajj et al., 2011 *(290)*	Cohort	Fertility/ ART	Bisulfite pyro-sequencing	PureSperm gradient (40 and 80%)	106	Age not detailed; subfertile males	28	Not detailed	Significant association between DNAm of *ALU* and outcome of ART	High

Rotondo et al., 2012 *(162)*	Case-control	Fertility/ ART	Bisulfite-PCR sequencing	PureSperm gradient (40 and 80%)	20	35 +/- 5.4; couples with recurrent spontaneous pregnancy loss	167	35.3 +/- 4.4; 20 males with proven fertility, 147 couples not affected by recurrent spontaneous pregnancy loss	*MTHFR* hypermethylation more common in semen from couples affected by recurrent spontaneous pregnancy loss	High
Ankolkar et al., 2012 *(291)*	Case-control	Fertility/ ART	Bisulfite Sanger sequencing	HiSep solution gradient	26	35.4; males with partners with recurrent spontaneous miscarriage	26	31.3; proven fertility	*H19* DNAm reduced in sperm from recurrent spontaneous miscarriage group	High
Camprubi et al., 2012 *(158)*	Cohort	Fertility/ ART	Bisulfite pyro-sequencing	Swim-up	107	36 +/-5.5 (26-53); subfertile males	30	26+/-6.15 (19-45); proven fertility	No significant association between rates of imprinting errors and outcome of ART	High
Kaminsky et al., 2012 *(292)*	Case-control	Neuro-logical disease	Bisulfite pyro-sequencing	Not detailed	43	40.2+/-11.6; 29 males with bipolar disorder 14 males with schizophrenia	30	37.7 (+/- 10.3); healthy volunteers	Hypomethylation of one *HCG9* CpG in sperm from individuals with bipolar disorder/schizophrenia	High
Kläver et al., 2012 *(293)*	Cohort	Method-ology	Bisulfite pyro-sequencing	Swim-up	10	Age not detailed; normozoospermic	N/A	N/A	No significant differences in DNAm in selected genes between cryopreserved and non-cryopreserved samples	High

Al-Khtib et al., 2012 *(160)*	Cohort	Fertility/ ART	Bisulfite pyro-sequencing	PureSperm gradient (50, 70 and 90%)	5	Age not detailed; proven fertility	None	N/A	Relative hypermethylation of *OCT4* and *NANOG* in sperm but hypomethylation in oocytes	Low
Berthaut at el, 2013 *(294)*	Case study	Medic-ations	Bisulfite pyro-sequencing	Ultra-centrifugation at 700g for 10 minutes	1	27; male undergoing temozolomide treatment. Nine samples in total	Not detailed	Age not detailed; proven fertility	Temozolomide treatment associated with lower sperm count (although remained normal) and *H19* hypomethylation	Very low
Rotondo et al., 2013 *(295)*	Case-control	Fertility/ ART	Methylation specific PCR and combined bisulfite restriction analysis	PureSperm gradient (40 and 80%)	10	36.1 (+/- 1.2); *MTHFR* promoter hypermethyl-ation. 5 with normal and 5 with abnormal semen parameters	10	Age not detailed; normal DNAm of the *MTHFR* promoter; 5 with normal and 5 with abnormal semen parameters	For normozoospermic samples, *H19* DNAm was associated with *MTHFR* DNAm	Moderate
Li et al., 2013 *(296)*	Case-control	Fertility/ ART	Methylation specific PCR (sequences analysed using BiQ Analyzer software)	Gradient (45, 60 and 90%) ultracentrifug ation. Both Percoll and PureSperm mentioned	40	Age not detailed; 20 oligozoo-spermic and 20 astheno-zoospermic men	20	Age not detailed; normo-zoospermic, proven fertility	No significant association between *H19* and *DAZL* DNAm and asthenozoospermia	Very low

Kläver et al., 2013 *(297)*	Cohort	Fertility/ ART	Bisulfite pyro-sequencing	Swim-up	212	Age not detailed; males under evaluation for subfertility	41	34.5 (+/- 7.2); 30.4 (+/- 6.0); 31 normo-zoospermic samples. Also 4-5 consecutive samples	*MEST* DNAm associated with oligozoospermia. DNAm levels at selected genes in normozoospermic males were stable for up to 951 days	Low
Montjean et al., 2013 *(298)*	Cohort	Fertility/ ART	Bisulfite targeted Sanger sequencing	Percoll gradient (45 and 90%)	175	Age not detailed; oligozoospermic	119	Age not detailed; normo-zoospermic	Abnormal DNAm of selected genes more prevalent in oligo-zoospermia. No association between DNAm levels and outcome of ART	Very low
Camprubi et al., 2013 *(299)*	Cross-sectional	Fertility/ ART	Bisulfite pyro-sequencing	Swim-up	6	Age not detailed; Subfertile males showing hypo-methylation of the *H19-ICR* locus; subset of a larger sample	None	N/A	No significant association between *CTCFL* mutations and *H19-ICR* sperm DNAm	Very low
Miao et al., 2014 *(300)*	Cross-sectional	Toxins	Methylation specific PCR	Not detailed	77	22-50; exposed to BPA	72	22-50; not exposed to BPA	Sperm *LINE-1* methylation level significantly lower in BPA exposed cases compared to that in the unexposed controls	Moderate

Botezatu et al., 2014 *(301)*	Case-control	Fertility/ ART	Methylation specific PCR	Not detailed	27	26-41; oligoasthenо-zoospermic	11	24-37; normo-zoospermic	Significant hypomethylation of *MTHFR* and *SNRPN* in samples with low sperm motility. No significant trend for other semen parameters	High/ Moderate
Richardson et al., 2014 *(302)*	Case-control	Fertility/ ART	Bisulfite pyro-sequencing	Swim-up	95	32-39; abnormal semen parameters (one or more of low count, low motility or abnormal morphology)	45	34-40; normo-zoospermic	The *RHOX* homeobox gene cluster is regulated by DNAm and *RHOX* gene cluster hypermethylation is significantly associated with semen abnormalities	Low
Kuhtz et al., 2014 *(303)*	Case-control	Fertility/ ART	Bisulfite sequencing	Swim-up and PureSperm gradient (40 and 80%)	27	Age not detailed; oligoasthenо-zoospermic	27	Age not detailed; normo-zoospermic	No significant differences in the occurrence of abnormal methylation imprints between sperm with and without morphological abnormalities	Very low
Guardiola et al., 2014 *(304)*	Cross-sectional	Tissue specificity	450K array	Not detailed	1	Age not detailed; publicly available methylome	N/A	N/A	*APOA1/C3/A4/A5* cluster is hypomethylated in liver but generally hypermethylated in other tissues analysed, including sperm	Low

Zhou et al., 2015 *(305)*	Case-control	Fertility/ ART	Methylation specific PCR and bisulfite sequencing PCR	Percoll gradient (50%)	48	30.13 (+/- 5.8); astheno-zoospermic	42	29.0 (+/- 4.8); normo-zoospermic	No DNAm of the *CRISP2* promoter; expression likely regulated by miR-27b	Moderate
Laurentino et al., 2015 *(306)*	Case-control	Fertility/ ART	Bisulfite pyro-sequencing and deep targeted bisulfite sequencing	Swim-up	26	34.0-39.33; males with abnormal sperm parameters	19	33.68 (+/- 1.58); proven fertility	Significantly increased variation in the DNA methylation values of the maternally methylated gene *KCNQ1OT1* in samples with abnormal sperm parameters	Moderate
Tian et al., 2015 *(307)*	Cross-sectional	Fertility/ ART	Methylation specific PCR	Samples washed in PBS and sperm wash buffer	118	31.4 (+/- 5.1); males under evaluation for subfertility	None	N/A	*LINE-1* negatively correlated with sperm motility	High
Montjean et al., 2015 *(308)*	Case-control	Fertility/ ART	Global DNAm assayed by immuno-staining for 5mC, targeted bisulfite sequencing	Percoll gradient	30	38.3 (+/- 6); Oligoastheno-zoospermic	62	38.5 (+/- 5.3); normo-zoospermic	Significant association between *H19* DNAm and sperm parameters	Low

Zhang et al., 2015 *(309)*	Cross-sectional	Cancer	Bisulfite pyro-sequencing	Not detailed	43	Age not detailed; adenocarcinoma of prostate	40	Age not detailed; benign prostatic hypertrophy	No significant difference in the detection of *RARB2* promoter methylation when comparing bisulfite pyro-sequencing in ejaculates to bisulfite sequencing of prostatic tissue	Very low
Soubry et al., 2016 *(310)*	Case-control	Diet/ lifestyle/ metabolic disease	Bisulfite pyro-sequencing	ISolate gradient (50 and 90%)	23	18-35; Males with an overweight/ obese BMI	44	18-35; lean males	Hypomethylation of *MEG3, NDN, SNRPN* and *SGCE/MEST0* DMRs in sperm from obese/ overweight males	High
Wu et al., 2016 *(311)*	Case-control	Method-ology	MethyLight	Not detailed	68	22-47; Azoospermic	24	22-47; normo-zoospermic	DNAm of the five selected testis-specific promoters was correlated between testicular DNA and paired cell free seminal DNA	Low
Zhang et al., 2016 *(312)*	Cross-sectional	Tissue specificity	Methylation specific PCR	Percoll gradient (40 and 80%)	15	23-34; oligoasthено-zoospermic	10	23-39; normo-zoospermic	Promoter DNAm correlates with tissue-specific expression of *BOULE* and *DAZL*	Very low
Xu et al., 2016 *(313)*	Case-control	Fertility/ ART	Methylation specific PCR	Not detailed	27	28; asthenozoo-spermic	25	28; normo-zoospermic	Asthenozoospermia significantly associated with abnormal DNAm of the *VDAC2* promoter	Low

Xu et al., 2016 *(313)*	Case-control	Fertility/ ART	Targeted DNAm quanti-fication using the MassArray Epityper	Percoll gradient (50%)	46	31.95 (+/- 3.1); asthenozoo-spermic	49	32.16 (+/- 3.26); normo-zoospermic	Lower levels of DNAm in *MEST, GNAS, FAM50B, H19, LINE-1* and *P16* in astheno-zoospermic males	Very low
Louie et al., 2016 *(314)*	Case-control	Fertility/ ART	Methylation -specific PCR	Swim-up for high concen-tration samples. Handpicked spermatozoa for low concentration samples	44	32.9-35.7; oligoastheno-zoospermic	Not detailed	34.1 (+/- 2.4); normo-zoospermic	No significant association between DNAm in selected imprinted genes, severe oligozoo-spermia and the *MTHFR* C677T SNP	Low
Atsem et al., 2016 *(315)*	Cohort	Aging	Bisulfite pyro-sequencing	Swim-up and PureSperm gradient (40 and 80%)	350	Age not detailed; 162 samples that had led to the birth of a child (cohort 1), 188 samples without pregnancy (replication cohort)	None	N/A	DNAm of *FOXK1* and *KCNA7* associated with paternal age in replication cohort	High

Buckley et al., 2016 *(316)*	Cross-sectional	Tissue specificity	Reduced represen-tation bisulfite sequencing (RRBS) and DNase I hyper-sensitive site profiles	Not detailed	ND	Age not detailed; publicly available methylomes	N/A	N/A	Testis and sperm hypomethylation of *BHMG1* and *RSPH6A* associated with testis-specific expression	Very low
Soubry et al., 2017 *(317)*	Cross-sectional	Toxins	Bisulfite pyro-sequencing	ISolate gradient (50 and 90%)	67	18-35; from TIEGER cohort	N/A	N/A	Mono-isopropylphenyl diphenyl phosphate associated with *MEG3, NDN, SNRPN* DNAm. Tris(1,3-dichloro-2-propyl) phosphate exposure associated with DNAm at *MEG3* and *H19*	High
Dong et al., 2017 *(318)*	Case-control	Fertility/ ART	Bisulfite pyro-sequencing	Percoll gradient (45 and 90%)	155	31.13-32.17; 48 oligozoospermic, 52 astheno-zoospermic, 55 teratozoospermic	50	32.22 (+/- 3.59); normo-zoospermic	In smokers, subfertility was associated with hypomethylation of *H19* in oligozoospermia and hypermethylation of SNRPN-ICR in azo- and teratozoospermia	Moderate /high

Kobayashi et al., 2017 *(319)*	Case-control	Fertility/ ART	Combined bisulfite restriction analysis	Swim-up	70	36.5 (+/-6.1); 40 males with moderate oligo-zoospermia and 30 males with severe oligo-zoospermia	151	35.4 (+/-5.4); normozoo-spermic	Oligozoospermia associated with smoking and high consumption of carbonated drinks. Adverse pregnancy outcomes associated with sperm DNAm errors	Moderate
Marques et al., 2017 *(320)*	Case-control	Fertility/ ART	Methylation specific PCR	Micro-manipulation from testicular biopsy	15	Age not detailed; astheno-zoospermic	Not detailed	Age not detailed; normozoo-spermic males unable to produce semen due to spinal cord injuries	No significant association between *H19* and *MEST* DNAm and asthenozoospermia	Very low
Nasri et al., 2017 *(321)*	Case-control	Fertility/ ART	Combined bisulfite restriction analysis	PureSperm gradient (40 and 80%)	23	Age not detailed; samples with below normal sperm parameters	11	Age not detailed; normo-zoospermic	No significant difference between *H19* DNAm and sperm parameters	Very low
Lu et al., 2018 *(322)*	Cross-sectional	Toxins	Bisulfite pyro-sequencing	Samples centrifuged at 200×g for 15 min. Sperm selection ND.	243	22-59; Exposed to varying levels of mercury	N/A	N/A	Negative correlation between sperm DNAm levels of *H19* and urinary Hg concentrations	High

Gunes et al., 2018 *(323)*	Case-control	Fertility/ ART	Methylation specific PCR	Somatic cell lysis	10	Age not detailed; oligoasthenо-zoospermic	29	Age not detailed; normozoo-spermic	In oligozoospermic men, *MLH1* was relatively hypermethylated. Seminal ROS levels were associated with *MLH1* DNAm	Very low
Tian et al., 2019 *(324)*	Cross-sectional	Toxins	High-melting resolution PCR	Not detailed	86	31.6; Males under evaluation for subfertility	N/A	N/A	High urinary levels of selected phthalate metabolites associated with decreased *LINE-1* promoter methylation and low sperm parameters	High
Darbandi et al., 2019 *(325)*	Case-control	Toxins	Methylation specific PCR	Centrifuged at 300g for 5 min. Sperm selection method ND	43	33-35; normozoo-spermic males with high ROS levels in seminal plasma	108	33-35; normozoo-spermic males with low to moderate ROS levels in seminal plasma	High ROS levels associated with DNAm of *H19* and *IGF2*, and with lower semen parameters	Moderate

Table 4-3. Candidate gene studies of DNA methylation in human sperm.
'Initial processing' refers to how the semen sample has been cleaned from seminal fluid, debris, somatic cells and immotile/dead spermatozoa. 'Gradient' refers to a discontinuous gradient ultracentrifugation using the medium detailed in the table. ND = Not Detailed, ROS = Reactive Oxygen Species

4.8.2 Genes studied in candidate gene analyses

As presented in Figure 4-2, the most common reasons for investigating DNA methylation of specific candidate genes was their involvement in imprinting, germ cell development and function, and embryonic development. Indeed, as presented in Figure 4-3, a relatively small number of genes were analysed in multiple studies.

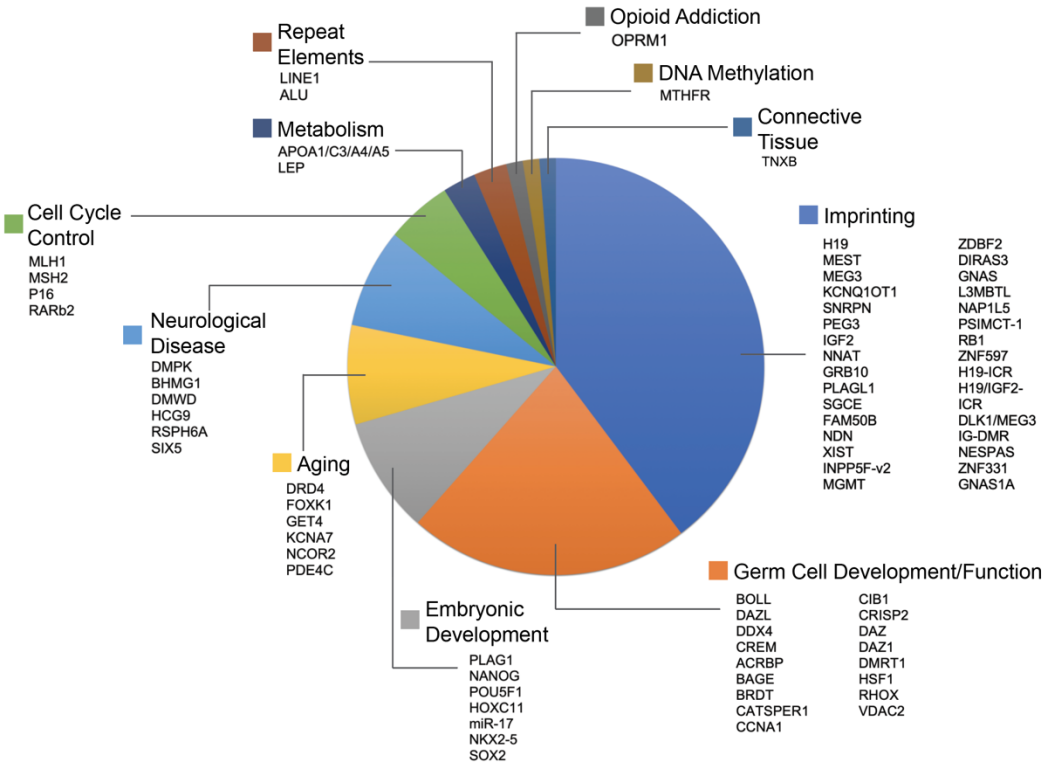


Figure 4-2. Overview of the basis for selection of candidate genes in studies of spermatozoal DNA methylation.
Most genes detailed in the figure have pleiotropic cellular functions. However, the biological processes described above are those used as a rationale for analysing their epigenetic regulation in spermatozoa specifically. Gene names used are those used by GeneCards (www.genecards.org).

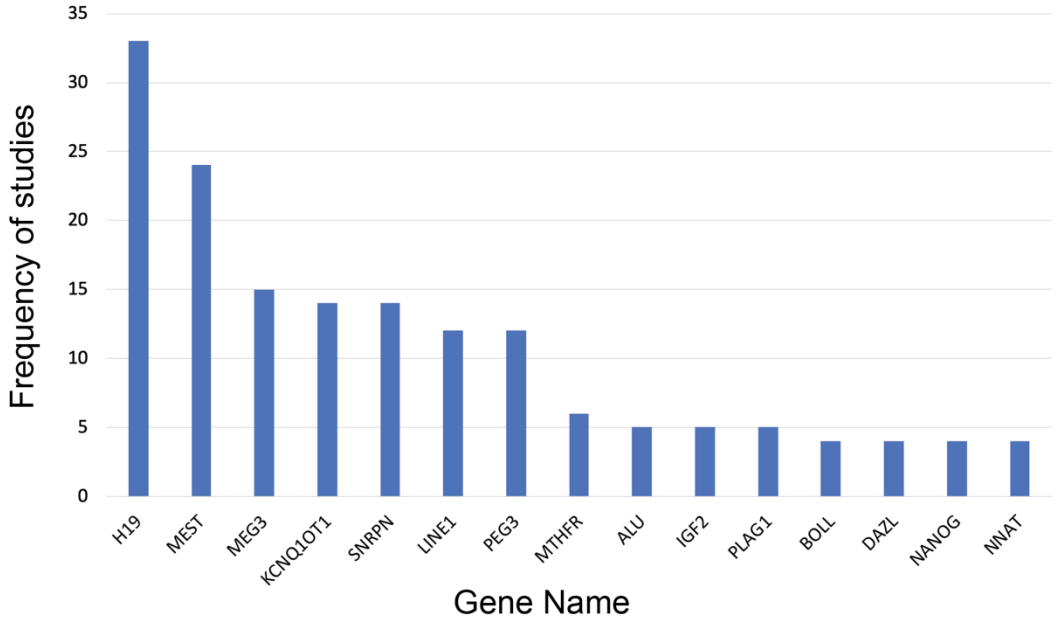


Figure 4-3. Frequency of studies investigating DNA methylation profiles of particular candidate genes.

4.8.3 Conclusions from studies of candidate genes in human sperm

The relative cost-effectiveness of candidate gene studies has made them an attractive method of studying a small number of genes posited to be involved in a particular biological process. However, these hypothesis-driven approaches do suffer from several methodological limitations.

First, our understanding of the biological pathway that the gene is part of is inevitably incomplete and may be wrong. Therefore, identifying an association between DNA methylation of a candidate gene and a particular phenotype does neither prove nor refute that the biological pathway that the gene is thought to be part of is disrupted. For example, low expression of a gene as a consequence of promoter DNA methylation may be compensated for by downstream effects of other pathway members.

Second, candidate gene analyses are prone to publication bias, especially if they are relatively small-scale. It is possible that several of the genes analysed in published studies of DNA methylation in sperm may have been investigated in similar studies

previously, but that small studies with negative results were less likely to become published.

Regarding generalisability of findings from candidate gene analyses, it is worth noting that at least two-thirds of studies (40 out of 61) recruited participants wholly or partially from a reproductive medicine setting. In a further five studies, it was unclear where participants had been recruited from.

4.8.4 Summary of findings from high quality studies of candidate gene analyses of DNA methylation

A couple of the high quality studies of candidate genes in human spermatozoa showed an association between abnormal semen parameters or poor outcomes of assisted reproductive technologies (ART) and abnormal DNA methylation in promoters of imprinted genes (278, 291). Similarly, a couple of studies showed an association between abnormal semen parameters or outcome of ART and levels of methylation at repetitive DNA elements (290, 307). In contrast, one high quality study showed that there was no association between rates of imprinting errors and the outcome of ART (158). Some high quality studies have shown an association between exposure to various toxins, such as organophosphates and mercury, and abnormal promoter methylation of imprinted genes or repetitive elements (317, 322, 324). One study showed that obesity could influence methylation of imprinted genes and genes associated with spermatogenesis (310). Given the methodological problems even with well-performed candidate gene analyses, such as potential publication bias and incomplete understanding of the biological pathways involved, it would be tentative to draw general conclusions from these sometimes conflicting results. Nevertheless, it is possible that relatively extreme situations of high toxin exposure, extreme obesity or severely abnormal sperm parameters are associated with abnormal sperm parameters and lower success rates of ART, but that these effects are negligible under more normal conditions. In the case of toxin exposure in particular, this would be in line with some of the high quality global analyses of spermatozoal DNA methylation discussed above.

Among the more replicated findings in high quality candidate gene analyses is hypermethylation of the methylenetetrahydrofolate reductase (*MTHFR*) promoter in association with subfertility, poor ART outcomes or toxin exposure (162, 281, 301). MTHFR catalyses a key step in folate metabolism, which is essential for the establishment of DNA methylation (326). Mutations in the *MTHFR* gene have reliably been linked to developmental perturbations related to folate metabolism, such as neural tube defects (326). It is possible, therefore, that hypermethylation of the *MTHFR* promoter, leading to its transcriptional repression, leads to inaccurate establishment of the sperm DNA methylome with downstream consequences such as poor outcome of ART.

One high quality study analysed CpG methylation of *HCG9* in sperm, post-mortem brain tissue and peripheral blood from individuals with schizophrenia and bipolar disorder (292). The study found that methylation at one CpG site was significantly different between cases and controls (292). To my knowledge, this finding has yet to be replicated and it should therefore be interpreted with caution. Nevertheless, it is interesting consider the now well-established association between paternal age and psychiatric disorders in his offspring, suggesting that there is an association between alterations to sperm during an individual's lifetime and the health of his offspring (327). In this regard, it is also interesting to note that age-related methylation changes were observed in *FOXK1* and *KCNA7* in another high quality study by Atsem et al (315).

The final high quality candidate gene study identified in this review showed that cryopreservation of sperm did not significantly alter the methylation status of selected genes (293). This is reassuring from a reproductive medicine point of view, as sperm is frequently cryopreserved for long periods of time in cases where, for example, individuals are due to undergo treatment that could negatively impact of their fertility and opt to preserve spermatozoa beforehand.

4.9 Genome-Wide Analyses of DNA Methylation in Human Sperm

Genome-wide analyses of human sperm methylation are presented in Table 4-4.

Author, year, reference	*Area of research*	*Method of assaying DNA methylation*	*Initial processing*	*N (cases)*	*Age (years); description of cases*	*N (controls)*	*Age (years); description of controls*	*Main Results*	*Study Quality*
Flanagan et al., 2006 (328)	Methodology	Methylation-specific PCR for DNAm measurement in selected genes and microarray-based genome-wide CpG profiling	ND (taken from cryo-storage)	46	Mean1 = 27 (22-35); Mean2 = 39 (24-56); healthy volunteers	N/A	N/A	Significant variation between DNAm levels of selected genes both within and between samples	High
Houshdaran et al., 2007 (329)	Fertility/ ART	MethyLight (bisulfite-based PCR) to measure DNAm levels in target genes. CpG methylation also assayed on a subset of samples using an Illumina bead array of 1421 CpG sites	ISolate gradient	65	22-49; partners of females undergoing fertility investigation	None	N/A	Association between sperm parameters and DNAm of *HRAS, NTF3, MT1A, PAX8, DIRAS3, PLAGL1, SFN, SAT2CHRM1* and *MEST*	Moderate
Rakyan et al., 2008 (330)	Methodology	MeDIP-Seq	ND	4	Age ND; normozoo-spermic	N/A	N/A	Generated a reference DNA methylome for human spermatozoa	Low
Fan et al., 2009 (331)	Methodology	Bisulfite sequencing of PCR amplicons (the Human Epigenome Project)	ND	1	24; from the Human Epigenome Project	N/A	N/A	CpG island methylation patterns more similar between somatic tissues than between the different somatic tissues and sperm respectively	Very low

Hammoud et al., 2009 (332)	Methodology	MeDIP followed by promoter arrays. Validation by bisulfite-sequencing of identified sites	Gradient ultracentri-fugation followed by somatic cell lysis	4	ND	N/A	N/A	Overall hypomethylation of developmental promoters compared to fibroblasts	Very low
Zeschnigk et al., 2009 (333)	Methodology	Restriction digestion of DNA followed by massive parallel bisulfite sequencing of CG-rich DNA fragments	ND	4	Age ND; pooled sperm samples	N/A	N/A	Identified 824 (13.7%) and 482 (8.5%) fully methylated autosomal CGIs in blood and sperm DNA respectively	Moderate
Molaro et al., 2011 (334)	Tissue specificity	Whole-genome bisulfite sequencing	ND	2	Age ND; healthy males	N/A	N/A	The features that determine DNA methylation patterns differ between male germ cells and somatic cells, and elements of these features have diverged between humans and chimpanzees	Low
Pacheco et al., 2011 (335)	Fertility/ ART	27K array	Percoll gradient (50%)	21	Age ND; males under evaluation for subfertility	None	N/A	Association between 9,189 CpGs and low motility	Moderate
Aston et al., 2012 (336)	Fertility/ ART	27K array followed by targeted bisulfite pyro-sequencing	ND	28	31.0+/-1.9; 15 males with ab-normal chroma-tin packaging + 13 males with poor embryo outcomes	15	37.1+/-2.0; normozoo-spermic	No significant CpG methylation differences identified	Low

Li et al., 2012 (337)	Genetics	Whole-genome bisulfite sequencing (WGBS)	ND	4	Age ND; healthy volunteers (2 from previously published research)	N/A	N/A	Association between regions of hypo-methylation and genomic instability	Moderate
Krausz et al., 2012 (168)	Tissue specificity	450K array	Percoll gradient (25%) and Swim-up	8	33-45; healthy volunteers	N/A		DNA methylation in normozoospermic men remains highly uniform regardless of the quality of sperm sub-populations	Moderate
Schütte et al., 2013 (338)	Fertility/ ART	27K array followed by targeted bisulfite pyro-sequencing	Swim-up procedure and PureSperm gradient (40 and 80%)	38	27-54; males under evaluation for subfertility (two samples each)	4	Age ND; normo-zoospermic	Unsupervised clust-ering analysis of 27k array data showed clustering by sperm count; no candidate genes showed signifi-cant difference	Moderate /Low
Kim et al., 2013 (339)	Tissue-specificity	27K array	ND	21	Age ND; publicly available methylomes	N/A	N/A	Significant hypo-methylation of *CTA* regulatory regions in cancer tissue and sperm compared to healthy somatic tissue	Very low
Jenkins et al., 2014 (340)	Aging	450K array and bisulfite pyro-sequencing of LINE (surrogate for global levels of DNAm)	No sperm selection method employed	17	23-56; proven fertility. Two samples taken 9-19 years apart	N/A	N/A	Identified 147 regions where DNAm was significantly associated with age	Moderate

Zeng et al., 2014 (341)	Tissue specificity	Whole-genome bisulfite sequencing (WGBS)	Silica-based gradient ultracentrifugation	2	Age ND; from previously published study; anonymous donors	N/A	N/A	CpG islands relatively hypomethylated in sperm compared to somatic tissues	Very low
Denham et al., 2015 (148)	Diet/lifestyle/ metabolic disease	450K array and 5mC ELISA assay kit	PureSperm gradient (40 and 80%)	12	24.4 (5.19); randomly assigned to exercise intervention	12	22.45 (4.75); randomly assigned to not exercise during a 3 month period	Global and genome-wide alterations in DNA methylation following 3 months of exercise	Moderate
Feinberg et al., 2015 (342)	Neurological disease	CHARM 3.0 array, including 30 samples also run on a 450K array	No sperm selection method employed	44	27-51.2; fathers in an cohort enriched for autistic children	None	N/A	193 DMRs in paternal sperm associated with performance on the Autism Observational Scale for Infants at 12 months of age in offspring	Low
Urdinguio et al., 2015 (343)	Fertility/ ART	450K array, targeted analyses by bisulfite pyrosequencing for selected genes (*LINE-1, Alu Yb8, NBL2, D4Z4*), global 5mC levels measured by immunostaining for 5mC	PureSperm gradient (65 and 90%)	7	30-55; 29 subfertile normozoospermic males (only 7 included in array analysis)	5	22-49; 17 normozoo-spermic fertile males (only 5 included in array analysis)	A number of CpG sites significantly differentially methylated between subfertile cases and fertile controls	Low
Jenkins et al., 2015 (344)	Methodology	450K array	ISolate gradient (35 and 90%)	20	35.69 (+/- 2.0); males under evaluation for subfertility	N/A	N/A	Higher DNAm variability in low quality sperm	Low

Aarabi et al., 2015 (345)	Diet/lifestyle/ metabolic disease	Reduced representation bisulfite sequencing (RRBS) and targeted bisulfite pyro-sequencing of *H19, DLK1/GTL2 IG-DMR, KCNQ1OT1, PLAGL1, MEST, SNRPN*	ND	30	37.9 (+/- 1.3); subfertile males, normozoo-spermic	None	N/A	No significant improvements in sperm parameters with folic acid supplementation	Low
Aston et al., 2015 (163)	Fertility/ ART	450K array	PureSperm gradient (45 and 90%). Some analyses performed on whole ejaculates	127	Age ND; 54 with successful pregnancy outcomes and 72 with poor embryogenesis	54	ND; normozoo-spermic, proven fertility	DNAm may be predictive of embryo quality, but not of IVF outcome	Very low
Donkin et al., 2016 (346)	Diet/lifestyle/ metabolic disease	Reduced representation bisulfite sequencing	Swim-up	16	24-40; 10 obese males and 6 obese males undergoing bariatric surgery	13	30-39; lean males	No CpG met genome-wide significance	Low
Jenkins et al., 2016 (165)	Fertility/ ART	450K array	Somatic cell lysis	29	27.55 (+/- 0.71); subfertile males	27	29.74 (+/- 0.71); proven fertility	DNAm of *HSPA1L* and *HSPA1B* significantly associated with pregnancy rates	Very low
Du et al., 2016 (347)	Fertility/ ART	Liquid hybridization (promoter capture) capture-based bisulfite sequencing	Percoll gradient (47.5, 57, 76 and 95%)	7	31.7 (+/- 4.0); asthenozoo-spermic	8	29.8 (+/- 3.2); normozoo-spermic	No significant association between DNAm or DNAm variability and asthenozoospermia	Low/very low

Jenkins et al., 2016 (348)	Fertility/ ART	450K array	ISolate gradient	94	32.0-36.32; classified according to semen parameters	N/A	N/A	No DNAm alterations associated with teratozoospermia	Low
Dere et al., 2016 (349)	Fertility/ ART	450K array	Somatic cell lysis	12	34 (+/- 7); 2 samples each	N/A	N/A	Intra-individual methylation levels between successive samples correlated with one another more strongly than inter-individual DNAm	Low
Camprubi et al., 2016 (164)	Fertility/ ART	450K array	Somatic cell lysis	42	38.36 (+/- 5.31); subfertile males	19	25.95 (+/- 4.80); proven fertility	Associations between age and DNAm of *RPS6KA2*; oligozoo-spermia and DNAm of *APCS*; chromosome abnormalities and DNAm of *JAM3/ NCAPD3* and between fecundity and DNAm of *ANK2*	High-Moderate
Wu et al., 2017 (350)	Toxins	450K array	Gradient (90%) ultracentri-fugation. Medium ND	48	18-55; males under evaluation for subfertility	None	N/A	131 DMRs were associated with at least one urinary phthalate metabolite	Moderate
Shnor-havorian et al., 2017 (351)	Medications	MeDIP-Seq. Validation of MeDIP-identified DMRs using minimal read depth bisulfite sequencing	ND	9	19.12-29.86; males with previous cisplatin-based treatment for osteosarcoma	9	27.5-44.4; healthy volunteers	A signature of significant DMRs identified in chemotherapy-exposed sperm	Very low

Fukuda et al., 2017 (352)	Tissue specificity	Whole-genome bisulfite sequencing (WGBS)	ND	4	25-30; publicly available methylomes	N/A	N/A	Sperm methylomes contained more hypomethylated domains than did the somatic methylomes	Very low
Camprubi et al., 2017 (353)	Fertility/ ART	450K array	Somatic cell lysis	19	25.95 (1.41); proven fertility	N/A	N/A	Promoter CpGs relatively hypomethylated in spermatozoa. Identified 94 genes that appear resistant to demethylation	Very low
Chan et al., 2017 (354)	Diet/ lifestyle/ metabolic disease	450K array, targeted DNAm quantification using the MassArray Epityper, restriction landmark genomic scanning, methyl-CpG immunoprecipitation followed by hybrid-isation to human CpG island microarrays	Ultracentri-fugation at 9400g for 20 minutes	18	26-36; 10 samples from men taking 400 microgram folate/day for 90 days. Additional cohort of 8 who had been exposed to folate-fortified food long term	9	33 (+/- 2); males taking placebo rather than folate	No significant association found	Very low
Jenkins et al., 2017 (355)	Toxins	450K array	Somatic cell lysis	78	32.4 (+/- 0.9); smokers	78	31.2 (+/- 0.6); non-smokers	Smoking associated with DNAm in 141 CpG sites	Moderate

Cheng et al., 2017 (356)	Fertility/ ART	Agilent custom 1M Promoter-CpG island microarray followed by targeted bisulfite pyro-sequencing	Enzymatic digestion and counter-current centrifugal elutriation	17	33.2 (+/- 0.5); 1 sample from a male with moderate hypo-spermiogenesis selected for whole genome DNAm profiling. Then selected candidate genes evaluated in 16 other males with HS	1	ND; 9 normozoo-spermic males with obstructive azoo-spermia, 1 of which was selected for whole genome methylation profiling	Hypospermiogenesis associated with DNAm of *BOLL, DDX4, HORMAD1*, and *MAEL*	Very low
Laqqan et al., 2017 (357)	Fertility/ ART	450K array followed by targeted bisulfite sequencing	Gradient (50%) ultracentrifu gation. Medium ND	70	35.89 (6.03); 15 for screening and 55 for validation	71	33.98 (5.53); 15 for screening and 56 for validation	Significant associations between subfertility and altered methylation levels in CpGs related to *ALS2CR12, ALDH3B2, PRICKLE2*, and *PTGIR*	High
Laqqan et al., 2017 (358)	Fertility/ ART	450K array followed by targeted bisulfite sequencing	Somatic cell lysis and PureSperm gradient (45 and 90%)	101	36.4 (3.24); 20 + 81 (screening + validation) oligospermic men	66	36.4 (3.24); 20 + 44 (screening + validation) males with proven fertility	Association between oligozoospermia and DNAm of *UBE2G2* and cg04807108	Moderate

Laqqan et al., 2017 (167)	Fertility/ ART	450K array followed by targeted bisulfite sequencing	Somatic cell lysis	72	37.4 (6.1); 15 + 57 (screening and validation) from subfertile men (unable to conceive for at least 10 years)	36	38.5 (5.2); 15 + 21 (screening and validation) men with proved fertility	Association between subfertility and DNAm in *KCNJ5, MLPH* and *SMC1β*	Moderate
Denomme et al., 2017 (359)	Fertility/ ART	450K array on a subset of sperm samples (12 out of 40). Beta value distribution examined for histone-retained regions	PureSperm gradient (45 and 90%), swim-up and somatic cell lysis	20	40.3 (+/- 1.3); normozoo-spermic samples (20 that led to 'good' embryogenesis)	20	42.0 (+/- 1.1); normozoo-spermic sample (20 that led to 'poor' embryo-genesis)	Significant association between DNAm in 1634 CpG sites and embryo quality	Low
Laqqan et al., 2017 (360)	Toxins	450K array followed by targeted bisulfite sequencing	Sperm separation medium gradient (45 and 90%) and somatic cell lysis buffer	36	40.39 (+/- 7.32); smokers (15 for screening, 36 for validation)	42	38.38 (+/- 8.27); non-smokers (15 for screening, 42 for validation)	Smoking associated with DNAm of *MAPK8IP* and *TKR*	High
Pilsner et al., 2018 (361)	Toxins	WGBS	ISolate gradient (50 and 90%)	4	18-19; males with a relatively high serum dioxin concentration	4	18-19; males with a relatively low serum dioxin concent-ration	Serum dioxin concentration associated with DNAm in 52 DMRs	Low

Ingerslev et al., 2018 (149)	Diet/ lifestyle/ metabolic disease	Reduced representation bisulfite sequencing	Swim-up	12	18-28; providing 3 samples each	None	N/A	330 DMRs after training and 303 DMRs after the detraining period	Moderate
Murphy et al., 2018 (362)	Toxins	Reduced representation bisulfite sequencing	ND	12	21.8 (3.8); cannabis users	12	25.8 (6.7); non-cannabis users	Cannabis use associated with DNAm at 3,979 CpG sites	Low
Jenkins et al., 2018 (363)	Aging	450K array data from previous studies. 148 regions previously identified to be strongly associated with the aging process in sperm to train age prediction model	ND	329	23-56; pooled data from previous studies. Mix of subfertile patients, sperm donors, and healthy volunteers	10	ND; 60 samples from 10 donors (validation cohort)	Model capable predicting age with an R2 of 0.89, a mean absolute error of 2.04 years, and a mean absolute percent error of 6.28%	Moderate
Al Khaled et al., 2018 (364)	Toxins	450K array followed by validation with bisulfite sequencing	PureSperm gradient (45 and 90%)	92	25-50; fertile smokers. 14 for discovery cohort and 78 for validation	14	25-50; fertile non-smokers	Smoking associated with DNAm in *PGAM5, PTPRN2* and *TYRO3*	Low
Laqqan et al., 2018 (166)	Fertility/ ART	450K array followed by targeted bisulfite sequencing	Somatic cell lysis	50	38.7 (6.9); subfertile males	28	38.5 (5.6); proven fertility	Association between subfertility and DNAm of *PRRC2A, ANXA2, MAPK8Ip3* and *GAA*	Moderate

Table 4-4. Genome-wide analyses of DNA methylation in human sperm

'Initial processing' refers to how the semen sample has been cleaned from seminal fluid, debris, somatic cells and immotile/dead spermatozoa. 'Gradient' refers to a discontinuous gradient ultracentrifugation using the medium detailed in the table. 450K array = Illumina HM450 Methylation array, COBRA = Combined bisulfite restriction analysis, ART = Assisted Reproductive Technologies, ND = Not Detailed.

4.9.1 General features of genome-wide analyses of human sperm

For the majority of studies (27/48), the quality of evidence was rated as low or very low. Among common reasons for downgrading studies were lack of phenotypic information about participants, lack of information on sample processing, lack of accounting for confounders such as age and smoking, lack of correcting for multiple testing and small sample sizes. Indeed, the sample sizes were generally lower than those in global analyses of DNA methylation and in candidate gene studies; the median number of total participants (cases and controls combined) was 24 (range 12-339), and the average was 48. The average was inflated by one study in particular that used pooled data from several previous studies to generate a total sample size of 339 (363). We decided to include both this large study and the smaller studies from which the data was pooled as they assessed different outcomes (e.g. the large study analysed age-related spermatozoal DNA methylation while the smaller studies investigated conditions such as subfertility and abnormal sperm parameters).

Studies where the quality of evidence was rated as high typically used multivariate analyses to account for covariates, or validated results in independent cohorts, such as in two studies by Laqqan *et al.* (167, 360).

The variability in sample sizes may help to explain why studies with similar methodologies have yielded contradictory results. For example, a case-control study Jenkins et al using the Illumina 450K methylation array (the HM450) that included 78 smokers and 78 non-smokers identified 141 CpG sites that were significantly associated with smoking (355). Al Khaled et al, however, did not identify any CpG sites that were significantly associated with smoking when using a similar protocol, which may be explained by including only 19 smokers and 20 non-smokers (355, 364). Also, results could have differed because of differences in length and amount of smoking among smoking participants in the two studies.

4.9.2 Summary of findings from high quality studies of genome-wide DNA methylation

Possibly the most striking conclusion from high quality studies of genome wide DNA methylation in human sperm is the lack of overlap between these findings and findings from candidate gene analyses. For example, Camprubi et al used the HM450 array to study genome-wide CpG methylation in subfertile cases and fertile controls, and also analysed results with regards to semen parameters and age (164). None of the intensively studied candidate genes were identified as differentially methylated between subfertile cases and fertile controls (164). Instead, significant associations were found between age and DNA methylation of *RPS6KA2*, between oligozoospermia and DNA methylation of *APCS* and between fecundity and DNA methylation of *ANK2* (164). Similarly, Laqqan et al used the HM450 to analyse genome wide CpG methylation in subfertile cases and fertile controls, and did not identify any of the previously studied candidate genes as associated with subfertility (357). Instead, this study identified CpGs related to *ALS2CR12, ALDH3B2, PRICKLE2*, and *PTGIR* as significantly differentially methylated in subfertile cases (357). What is also evident from these two studies is that there is little overlap of results between them, even though the research methodology was similar. This is possibly due to the relatively small sample sizes compared to EWASs of more accessible somatic tissues.

Other high quality studies include one that analysed genome wide CpG methylation in smokers versus non-smokers (360). The study identified DNA methylation at *MAPK8IP* and *TKR* as significantly associated with smoking (360). Given that paternal smoking has been consistently linked to birth defects and poor intrauterine growth of his offspring, further studies like these are warranted (227). Also, there is a well-established relationship between paternal aging and the risk for psychiatric and other conditions in his offspring (227, 327). It is therefore interesting that a genome-wide analysis of sperm CpG methylation identified age-related methylation changes in a number of genes, including *EED, CTNNA2, CALM1, CDH13* and *STMN2* (328). Notably, none of the identified genes overlapped with those studied in candidate gene analyses of age related changes in DNA methylation (315).

4.10 Recommendations for Future Research

4.10.1 Studies of global DNA methylation

Studies investigating global DNA methylation in human sperm have yielded inconclusive results, and commonly suffered from methodological problems. Importantly, even a well-powered and methodologically adequate study of global DNA methylation would, by definition, not yield information on specific genes or biological pathways that may be epigenetically perturbed in particular phenotypes. Therefore, the value of global DNA methylation analyses to understand the influence of DNA methylation on conditions such as subfertility or the potential for the sperm methylome to respond to environmental change is limited. It would be advisable to focus efforts on genome-wide approaches with an improved potential for uncovering biologically relevant findings.

4.10.2 Candidate gene studies

Candidate gene analyses suffer from methodological drawbacks as discussed above. They also, by definition, only investigate known genes or known regulatory regions of the genome, and thus leave the vast majority of the genome unexplored. Results from candidate gene analyses thus far have been inconclusive, and few findings have been replicated in genome-wide analyses. It would be recommendable to shift the focus towards unbiased, genome-wide approaches in future investigations.

4.10.3 Genome-wide studies

Unbiased, genome-wide CpG methylation studies of biologically relevant tissues have yielded important insights in a range of common human conditions, including metabolic and psychiatric disease (106, 365). Using this unbiased approach to investigate the sperm methylome holds the most promise in terms of improving understanding of conditions such as subfertility and the impact of environmental conditions on germ cells.

The majority of genome-wide investigations have employed the Illumina HM450 array, which captures around 1.7% of the ~ 28 million CpG sites in the genome.

Notably, the array only covers a small proportion of enhancers and other regulatory regions. Future investigations using the novel version of this array, the MethylationEPIC array, will generate an improved coverage of such regulatory regions with ~ 350, 000 additional CpG sites in enhancers (199).

It is evident that for robust, reproducible findings to be achievable, sample sizes need to be increased. Performing a power calculation prior to study commencement is recommended, for example by using published guidance for the MethylationEPIC array (366). As sperm is not routinely collected outside a reproductive medicine setting, it is likely that multicentre collaborations will be required to meet the requirement for increased sample sizes. Such collaborations should harmonise protocols for phenotyping of participants, processing samples and perform methylation analyses in order to limit batch effects. Until whole genome bisulfite sequencing becomes a cost-effective possibility for large-scale investigations, microarray-based approaches will likely be the best option. In addition, results from these arrays can be compared across studies. For example, the MethylationEPIC covers >90% of CpG sites present of the HM450 array (199). Therefore, results obtained from analyses using the MethylationEPIC can be checked for consistency with results from studies using the HM450 provided that similar methodologies were used.

Several genome-wide investigations thus far have been microarray-based analyses of subfertile cases and fertile controls (164-167, 343). Although these studies individually have identified a number of CpG sites that seem to be differentially methylated between cases and controls, there has been little overlap between study findings. It would be interesting to see if these identified sites are replicated in future, larger-scale investigations. It would also be advisable to combine DNA methylation analyses with functional studies to better characterise the biological pathways involved and their potential relevance to particular phenotypes.

4.10.4 Overall considerations

Future research should present detailed phenotyping of participants, including information on matching of participants in case-control studies as well as exclusion criteria for study participation. Factors to consider include age, smoking status, medication use, BMI, abstinence, drug/alcohol use, diet and sample storage time and conditions. In future studies, multivariate regression analyses that take potentially confounding factors into account would generate more reliable results.

The vast majority of studies performed to date have been cross-sectional. Longitudinal cohort studies with serial sampling would have a greater potential to improve our understanding of disease pathogenesis and identify biomarkers. Together with functional analyses, such studies would also begin to elucidate causality, i.e. whether particular sperm DNA methylation signatures are a cause or an effect of a particular phenotype.

There is a range of protocols for selecting motile spermatozoa and cleaning semen samples from contaminating somatic cells. Of importance is that irrespective of which protocol is employed, samples should be checked post-processing to ensure that spermatozoa used for epigenetic analyses contain a pure sperm population. Bioinformatically assessing for somatic cell contamination would also be advisable.

For studies of tissue specificity, e.g. comparing the DNA methylome of a somatic tissue to that of spermatozoa, samples should be matched. Several studies included in this review compared the DNA methylome of spermatozoa to that of somatic tissues of different individuals. Because genetic variation influences the DNA methylome, not using matched tissues increases confounding and should be avoided in future research.

The generalisability of findings to the general male population from the majority of studies included in this review is debatable, as they recruited participants from reproductive medicine settings. This selection bias would be avoided by recruiting study participants that are more representative of the male population as a whole. It

is likely that multicentre collaborative efforts would be required to generate large enough sample sizes to meet this aim.

The majority of studies have focused on differential DNA methylation related to subfertility or abnormal semen parameters. Widening the scope in future investigations to include more research on, for example, the influence of metabolic health, toxins and aging on sperm epigenetics would be recommended. Such research would also be more likely to help explain the epidemiological associations between paternal aging, smoking and BMI on offspring health (227). In this regard, however, other causes than changes to spermatozoa, whether genetic or epigenetic, need also to be taken into account. For example, there has long been an assumption that increased paternal age is associated with an increased risk of conditions such as autism and schizophrenia in his offspring due to age-related increases in the rate of *de novo* mutations in spermatozoal genomes (367). However, recent research suggests that a more influential part of the explanation for the association between paternal age and offspring psychiatric illness is that psychiatric illness is associated with an increased propensity for men to father his first child either at a relatively young or at a relatively old age (367). These two explanations for the association are of course not mutually exclusive, and it is possible that the increased rate of *de novo* mutations also play a part. Nevertheless, future research that explores the association between paternal age-related changes to his (epi)genome and the propensity for his offspring to develop psychiatric disease should take these behavioural or social factors into account.

This systematic review has focused exclusively on DNA methylation as opposed to other epigenetic mechanisms, partly because DNA methylation has been the most widely studied epigenetic mechanism. However, other epigenetic mechanisms such as small RNA species and histone modifications may also influence the sperm epigenome and effects between generations and therefore require study (144, 145). Our understanding of crosstalk between the nucleotide sequence, the DNA methylome, histone modifications and non-coding RNA is also in its infancy. Future research that elucidates the integration of different epigenetic modifications would

significantly enhance our understanding of the dynamic nature of gene expression in spermatozoa.

4.11 Strengths and Limitations of the Systematic Review

The major strength of this systematic review is the comprehensive literature search identifying a large number of relevant publications. It is nevertheless possible that there are relevant publications which fit the scope of this review, but which were not identified by our search terms.

We needed to modify the GRADE criteria to fit with the particular purpose of this review (244). It is possible that we should have included different rating criteria, for example modifying our assessment of studies depending on whether they were of a case-control EWAS type or whether they were studies of tissue specificity of DNA methylation. Also, the heterogeneity of studies included in this systematic review prevented us from adding meta-analyses where relevant. Meta-analyses might have given a better overview for the reader, and made summarising results more conclusive.

Further, methods of epigenomic investigation have been developing rapidly since 2003, i.e. the year from which studies were included in this review. This means that early studies may have been of relatively high quality at the time when they were performed, but rated as low quality in our review because they are compared to newer studies using improved methodologies.

Another limitation is that it is difficult to assess the risk of, for example, publication bias. There can be a suspicion that small-scale, low-cost analyses with negative results are less likely to be published than small-scale, low-cost analyses with positive results. However, it is by definition impossible to know what the ratio between these outcomes are and how such biases may have impacted the understanding of a particular field. For example, if all candidate gene analyses of subfertility with negative results had been published in addition to those with positive results, would

there have been a general understanding that subfertility is associated with abnormal methylation of imprinted genes?

4.12 Overall Conclusions

Detailed characterisation of the human sperm methylome has the potential to improve our understanding of subfertility, embryogenesis and the ability for environmental and acquired phenotypes to influence the next generation. From human studies performed to date, however, there are few, if any, robust and replicable findings that have significantly improved this understanding.

Importantly, in the context of intergenerational effects, there is yet no evidence of specific DNA methylation signatures in human spermatozoa that are able to influence offspring development. A large number of studies have investigated imprinted genes in sperm under the presumption that imprinting abnormalities identified in spermatozoa harbour the potential to influence the physiology of the next generation. However, such studies fail to take into account the widespread resetting of the epigenome that occurs between generations. It would be interesting to investigate whether there is overlap between CpG sites identified as differentially methylated in particular phenotypes and the so-called 'escape variants', i.e. sites that appear resistant to the demethylation processes that occur after fertilisation and during gonadal development (128).

Lastly, it will be a long time before studies of DNA methylation in human spermatozoa reach a comparable magnitude to those that have been performed on peripheral blood. As an intermediate step, therefore, it would be interesting to characterise the epigenetic covariation between these tissues in a genome-wide, unbiased manner on matched sperm and blood samples. Such analyses would help to elucidate under which conditions, if any, peripheral blood could be used as a surrogate tissue for spermatozoa. It could also generate information on which specific CpG sites are highly correlated versus uncorrelated between the two tissues. The largest study to date that has performed this characterisation included a total of 8 participants, thus a more large-scale investigation is warranted (168).

Chapter 5
Covariation of DNA Methylation in Blood and Semen

5.1 Introduction

Studies in humans and non-human mammals suggest that paternal metabolic health could influence the growth of his offspring (68, 69, 127). Animal studies suggest that the association between paternal metabolic disease and fetal growth could be mediated through epigenetic changes to spermatozoa that are passed from a father to his offspring and affect metabolism in the developing fetus (140, 141). Such studies are lacking in humans.

Sperm represents the most relevant tissue to study in the context of whether acquired paternal traits could biologically impact on the next generation. However, human semen samples are not regularly collected and analysed outside of a reproductive medicine setting. They are also less readily obtainable than, for example, peripheral blood. Thus, the human sperm epigenome remains relatively poorly characterised compared to many somatic tissues.

Some studies have used human blood as a proxy tissue for studying epigenetic profiles in germ cells (153). However, such studies are problematic. First, epigenetic signatures are highly tissue- and cell- type specific (77). Secondly, they fail to address the widespread two-stage process of DNA demethylation that occurs shortly after fertilisation and during gonadal formation (128).

As discussed in detail in Chapter 4, previous analyses of DNA methylation in human sperm have focussed largely on studies of DNA methylation in subfertility, and there has been a lack of robust, reproducible results across studies. In addition, participants have generally been recruited from reproductive medicine settings, potentially making results and conclusions less applicable to the general population.

It will be a long time before DNA methylation studies of human germ cells reach comparable magnitudes to those performed on peripheral blood. As an intermediate step, therefore, there is value in characterising the DNA methylation covariation between these tissues in order to better appreciate under which circumstances blood could be used as a surrogate tissue for sperm. Although absolute levels of DNA

methylation are likely to be highly tissue specific, sites where methylation levels co-vary would be candidates for inferring spermatozoal DNA methylation from analyses of blood (368).

To this end, I characterised the DNA methylomes of matched human sperm and blood samples from healthy males of proven fertility in an unbiased, genome-wide manner using the Illumina MethylationEPIC Array. The study aimed both to characterise the DNA methylome of the two tissues in detail, and to identify sites where methylation levels co-vary between the two tissues. In addition, a cohort of obese, fertile males were included to identify obesity specific co-variation between sperm and blood. Lastly, DNA methylation data from spermatozoa was compared to that of almost 6,000 tissue samples from the Gene Expression Omnibus (GEO) database in order to identify CpG sites that are hyper- and hypomethylated in sperm relative to somatic tissues.

5.1.1 Declarations

The samples included in this study were collected by myself and a research midwife (Anna Greco). I performed DNA extraction of semen samples, as well as of blood samples in conjunction with Anna Greco. Bisulfite conversion of DNA from blood and sperm samples was performed by myself and Ama Brew, research technician at the Blizard Institute, QMUL. The Illumina MethylationEPIC arrays were processed by Yasmin Panchbhaya at UCL Genomics, Great Ormond Street Institute of Child Health. Methylation data preprocessing and analysis was performed by Dr Sarah Marzi at the Blizard Institute, QMUL. The GEO analysis was performed by Dr Tyler Gorrie-Stone at the University of Essex.

5.2 Hypotheses

1) A. The human sperm epigenome displays large-scale overall differences compared with matched somatic tissues such as blood

   B. At a select number of CpG sites, there is a high DNA methylation correlation between the tissues

2) Obesity influences the sperm DNA methylome

5.3 Specific Objectives

1) To characterise the DNA methylomes of matched human sperm and blood samples from a cohort of healthy, fertile males in an unbiased, genome-wide manner using the Illumina MethylationEPIC Array

2) To identify CpG sites where DNA methylation levels co-vary between sperm and blood

3) To identify whether CpG sites that covary between sperm and blood in obese or lean males exclusively

4) To compare the DNA methylome of spermatozoa to that of methylation data from ~6,000 somatic tissues available on GEO

5.4 Methods

5.4.1 Study Outline

Participants were recruited as part of the Dad's Health Study at University College London Hospital (UCLH) between May 2016 and March 2019 as described in detail in Chapter 2 sections 2.3-2.8. All participants provided written, informed consent. Ethical approval was granted from the South East Coast - Surrey Research Ethics Committee on 28 September 2015 (REC reference number 15/LO/1437, IRAS project ID 164459). The study was also registered with the UCLH Joint Research Office (Project ID 15/0548).

Participants were recruited into three cohorts; 1) a discovery cohort consisting of 48 lean, fertile males (BMI 18-25 kg/m^2) who provided one sample of blood and one sample of sperm each, 2) a replication cohort consisting of 24 lean, fertile males (BMI 18-25 kg/m^2) who provided one sample of blood and one sample of sperm each, and 3) a cohort of overweight and obese, fertile males (BMI >25 kg/m^2) who provided one

sample of blood and one sample of sperm each. The cohort of overweight/obese males is hereafter referred to as the obesity cohort.

Participants in the discovery, replication and obesity cohorts each completed a questionnaire enquiring about past medical, family and treatment history. Participants were phenotyped with regards to height, weight, waist circumference and blood pressure by a trained research doctor or research midwife.

DNA was extracted from blood and sperm shortly after collection as described in section 5.4.6 below and in detail in Chapter 2 section 2.10.5. Extracted DNA underwent bisulfite conversion as described in Chapter 2 section 2.11. Bisulfite converted DNA samples were analysed for genome-wide CpG methylation levels using the Illumina MethylationEPIC Array at UCL Genomics, Great Ormond Street Institute of Child Health, as described in Chapter 2 sections 2.12. The output of the Illumina MethylationEPIC array analysis was provided as an intensity data (IDAT) file together with information on sample runs and chip positions.

A series of quality control and preprocessing steps, as described in Chapter 2 section 2.12 were applied to the DNA methylation data before analysis of CpG methylation with regards to the specific aims of the study was interrogated. An overview of the quality control, data preprocessing and analysis pipeline provided in Figure 5-1.

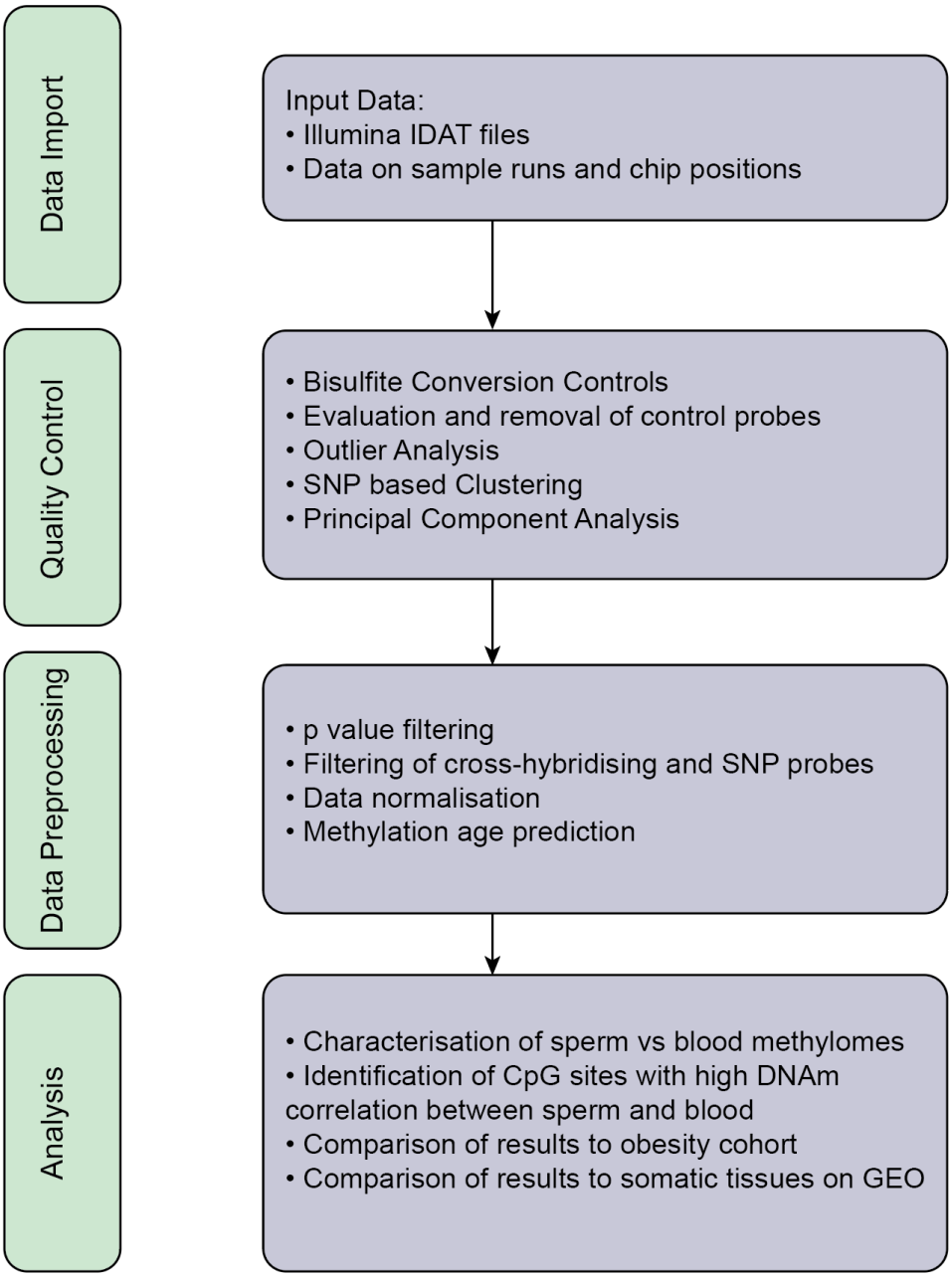


Figure 5-1. Overview of the analysis pipeline for analysing DNA methylation covariation in blood and sperm.
IDAT = intensity data, SNP = single nucleotide polymorphism, DNAm = DNA methylation.

5.4.2 Study Population

Participants included in the discovery and replication cohorts were recruited according to the following criteria:

- Aged 18 to 50 years
- No significant medical problems (as self-reported in study questionnaires)
- No regular use of medications (as self-reported in study questionnaires)
- No drug, alcohol or substance abuse
- Proven fertility
- BMI 18-25 kg/m^2

Participants included in the obesity cohort were recruited according to the following criteria:

- Aged 18 to 50 years
- No significant medical problems (as self-reported in study questionnaires)
- No regular use of medications (as self-reported in study questionnaires)
- No drug, alcohol or substance abuse
- Proven fertility
- BMI >25 kg/m^2

5.4.3 Statistical Analyses

Phenotype analyses were carried out using RStudio version 1.1.456. Results were independently verified by Dr Aviva Petrie at the UCL Eastman Dental Institute Biostatistics Unit using STATA 15 (StataCorp LLC, Texas USA).

Methylation data was processed and analysed using the wateRmelon package in R (203). The GEO analyses were performed using the bigmelon package in R (211).

5.4.4 Sample Collection, Handling and Processing

Detailed information regarding sample collection, handling of processing of samples collected as part of the Dad's Health Study can be found in Chapter 2 section 2.8.

Brief descriptions of how blood and semen samples included in the study of DNA methylation covariation were processed prior to epigenetic analysis is provided below.

5.4.5 Collection and Processing of Blood Samples

Peripheral blood samples were taken from the median cubital vein and clinical biochemistry analyses were performed by the UCLH clinical biochemistry laboratory. Blood to be used for DNA extraction was centrifuged at 3000g for 15 minutes within one hour of venepuncture. Buffy coat was isolated and used for DNA extraction. DNA was extracted using the Qiagen QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany) as described in Chapter 2 section 2.8.1 and stored in -80° prior to analysis. Bisulfite conversion of 500 (±10%) ng DNA derived from blood samples was performed using the D5001 EZ DNA Methylation Kit (Zymo Research, Irvine, CA, USA) according to manufacturer's instructions as described in Chapter 2 section 2.11. Bisulfite converted samples of DNA were stored at -20° for a limited time (<3 days) prior to being randomised and transported to UCL Genomics at the Great Ormond Street Institute of Child Health for immediate analysis using the Illumina MethylationEPIC Array.

5.4.6 Collection and Processing of Semen Samples

Semen samples were produced in participants' own homes or in designated rooms in the UCLH Reproductive Medicine Laboratory according to preference. Samples were obtained within one hour of sample production and initial processing of samples was performed in the UCLH Fertility and Reproductive Medicine Laboratory. First, samples were placed on a warm plate (35-37°C) for 20-30 minutes in order to liquefy. Sample volume was measured and 5µL of the sample was pipetted onto a Leja disposable counting chamber and analysed using the Computer-Assisted Sperm Analysis (CASA)/Sperminator software (Pro-Creative Diagnostics, Staffordshire, UK). Semen sample parameters measured were sperm concentration (millions/mL), percentages of sperm in four categories of motility from most motile to least motile (A – D), and average motile speed. Samples underwent gradient centrifugation (45 and 90% PureSperm medium) to select for the most motile sperm and to clean the

samples from somatic cells, seminal fluid and debris. DNA was extracted from the motile proportion of the sperm as described in section 2.10.5. Semen DNA was then stored in -80° prior to further processing and analysis. Bisulfite conversion of 500 (±10%) ng DNA derived from semen samples was performed using the D5001 EZ DNA Methylation Kit (Zymo Research, Irvine, CA, USA) according to manufacturer's instructions as described in Chapter 2 section 2.11. Bisulfite converted samples of DNA were stored at -20° for a limited time (<3 days) prior to being randomised and transported to UCL Genomics at the Great Ormond Street Institute of Child Health for immediate analysis using the Illumina MethylationEPIC Array.

5.4.7 Phenotype Characteristics of the Study Population

Phenotype characteristics of participants in the discovery, replication and obesity cohorts are presented in Table 5-1.

Measurements of BMI, waist circumference, systolic and diastolic blood pressure, total cholesterol as well as HDL and LDL cholesterol were found to be approximately normally distributed (Shapiro-Wilk test $p>0.10$) in all three cohorts, and are thus summarised as mean values and standard errors of the mean. The means of these measurements were compared between the discovery, replication and obesity cohorts using a one-way analysis of variance (ANOVA). This was followed by Tukey's Honest Significant Difference test if $p < 0.05$. The distribution of measurements of fasting glucose, C-reactive protein and triglyceride levels as well as calculated HOMA-IR values were found to be skewed (Shapiro-Wilk test $p<0.10$), and are therefore summarised as median values and interquartile ranges. In these cases, the Kruskal-Wallis test of ranks was used to compare median values between the cohorts. This was followed by Dunn's test of multiple comparisons using rank sums if $p < 0.05$.

As presented in Table 5-1, phenotype characteristics between the discovery and replication cohorts did not differ significantly (at the 0.05 level) for any of the studied phenotypic traits apart from in fasting glucose levels, where there was a weak but significant difference such that the discovery cohort had, on average, slightly higher values of fasting glucose ($p = 0.003$). However, because insulin resistance as

measured by HOMA-IR did not differ significantly between the discovery and replication cohorts, it is unlikely that the small difference in fasting glucose measurements would indicate a substantial difference in metabolic health between the two cohorts. It was thus determined that the discovery and replication cohorts were sufficiently matched in terms of phenotypic characteristics.

Chronological age is known to be associated with DNA methylation across several sites in the human genome (196). We therefore confirmed that there were no significant differences in age between the discovery, replication and obesity cohorts. There were, however, significant differences in BMI, waist circumference, fasting insulin levels, HOMA-IR and CRP between the discovery and obesity cohorts as well as between the replication and obesity cohorts (Table 5-1). The difference in BMI measurements is further detailed in Figure 5-2. In summary, there were substantial differences in metabolic profiles between the cohorts of lean, fertile males and the cohort of overweight/obese males respectively.

	Discovery	Replication	Obesity	p (difference between cohorts)	p (Discovery vs Replication)	p (Discovery vs Obesity)	p (Replication vs Obesity)
Age (years). Mean (SD)	36.3 (5.2)	34.1 (4.6)	35.1 (4.1)	0.192			
BMI (kg/m2). Mean (SD)	23.4 (4.6)	22.3 (1.1)	29.1 (3.2)	<0.001	0.060	<0.001	<0.001
Waist circumference (cm). Mean (SD)	84.4 (4.8)	82.4 (6.4)	99.4 (8.7)	<0.001	0.436	<0.001	<0.001
SPB (mmHg), average of two measurements. Mean (SD)	119 (11)	121 (10)	126 (9)	0.052			
DPB (mmHg), average of two measurements. Mean (SD)	77 (8)	78 (6)	81 (8)	0.050			
Total cholesterol (mmol/L). Mean (SD)	4.7 (0.7)	4.9 (0.9)	4.9 (1)	0.614			
HDL cholesterol (mmol/L). Mean (SD)	1.6 (0.3)	1.5 (0.3)	1.4 (0.3)	0.060			
LDL cholesterol (mmol/L). Mean (SD)	2.7 (0.7)	2.9 (0.8)	2.9 (0.9)	0.330			
Fasting glucose (mmol/L). Median (IQR)	4.8 (0.5)	4.6 (0.4)	4.7 (0.6)	0.018	0.003	0.088	0.105
Fasting insulin (mIU/L). Median (IQR)	5.3 (3.4)	5.1 (3.0)	8.9 (7.2)	0.002	0.309	<0.001	0.004
HOMA-IR. Median (IQR)	1.2 (0.8)	1.1 (0.6)	1.9 (1.4)	<0.001	0.285	<0.001	0.005
HOMA2-IR. Median (IQR)	1.1 (0.5)	0.6 (0.4)	1.1 (0.9)	0.014	0.048	0.414	0.003
CRP (mg/L). Median (IQR)	0.6 (0.3)	0.6 (0.1)	1 (1.8)	<0.001	0.105	0.001	<0.001
Triglycerides (mmol/L). Median (IQR)	0.9 (0.5)	0.9 (0.7)	1.2 (0.6)	0.282	0.335	0.056	0.157

Table 5-1. Phenotype characteristics of participants included in the discovery, replication and obesity cohorts
Legend continued overleaf

Table 5.1. Phenotype characteristics of participants included in the discovery, replication and obesity cohorts. *Legend continued from previous page:*
Reference ranges are derived from the UCLH Clinical Biochemistry Test Information sheet available from (230). The reference range for HOMA-IR is derived from (231). SD = Standard Deviation, IQR = interquartile range, BMI = Body Mass Index, SBP = Systolic Blood Pressure, DBP = Diastolic Blood Pressure, HOMA-IR = Homeostatic Model Assessment of Insulin Resistance, CRP = C-Reactive Protein, HDL = High Density Lipoprotein, LDL = Low Density Lipoprotein.

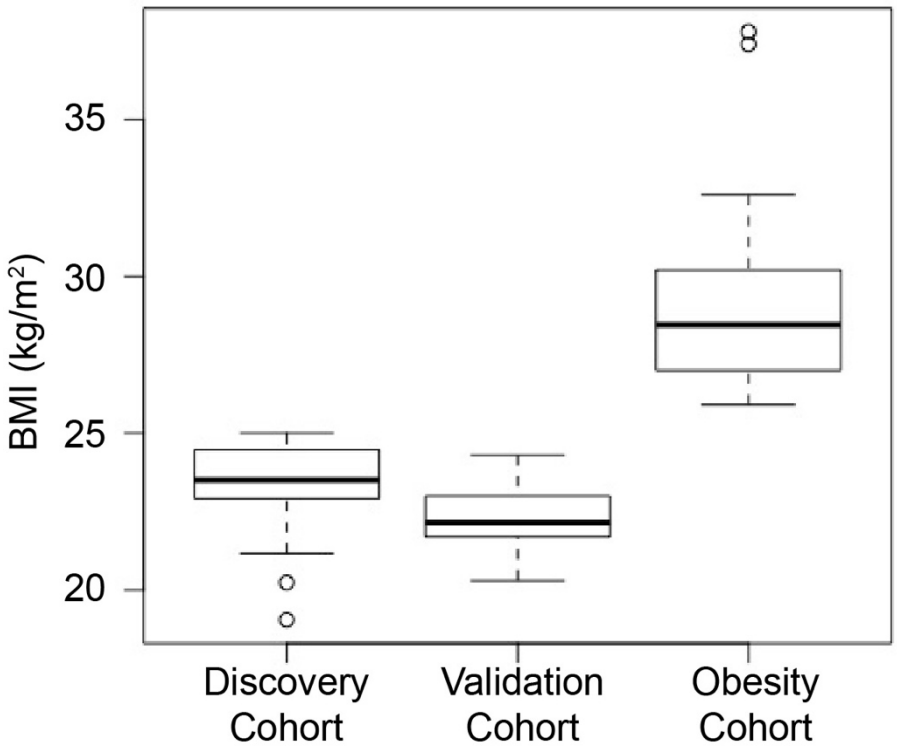


Figure 5-2. Comparison of BMI (kg/m2) between participants in the discovery, replication and obesity cohorts

5.4.8 Semen Sample Parameters of the Study Population

Semen parameters, of participants in the discovery, replication and obesity cohorts are presented in Table 5-2. As shown, semen sample parameters did not differ significantly between participants in the three cohorts.

	Discovery	Replication	Obesity	Reference Range	p
Volume (sperm, mL). Mean (SD)	2.9 (1.1)	2.9 (1.4)	2.6 (1.5)	> 1.5 mL	0.538
Concentration (sperm, millions). Mean (SD)	55.4 (37.2)	47.9 (33.9)	57.4 (31)	> 15 millions/mL	0.608
Total count per ejaculate (millions). Mean (SD)	161 (150.4)	149 (140.5)	157 (131.5)	> 39 million	0.953
Percentage A sperm. Mean (SD)	14.8 (10.6)	15.4 (10.6)	17.4 (10.3)	N/A	0.610
Percentage B sperm. Mean (SD)	23.9 (9.5)	22.1 (9.4)	20.4 (8.6)	N/A	0.348
Percentage C sperm. Mean (SD)	12.1 (3.7)	11.4 (3.7)	11.1 (4.3)	N/A	0.589
Percentage D sperm. Mean (SD)	49.3 (18.1)	50.5 (18.7)	51.1 (18.8)	N/A	0.926
Average motile speed. Mean (SD)	18.6 (2.6)	19.2 (4.4)	19.4 (2.3)	N/A	0.603

Table 5-2. Comparison of semen parameters between the discovery, replication and obesity cohorts

Semen sample parameters were measured using the Computer-Assisted Sperm Analysis (CASA)/Sperminator software (Pro-Creative Diagnostics, Staffordshire, UK). V = volume, C = concentration, SD = Standard Deviation, WHO = World Health Organization. Percentage A-D sperm refer to the proportion of spermatozoa in different motility grades where A = most motile and D = least motile. Reference ranges are derived from (245)

5.4.9 Results from the Discovery Cohort Quality Control: Bisulfite Conversion Efficiency and Evaluation of Control Probes

A quality control report from UCL Genomics, where the MethylationEPIC arrays were processed, reported high performance of the sample dependent and sample independent quality control probes discussed in Chapter 2 section 2.12.

Sample dependent controls include bisulfite conversion control probes and probes that assess the specificity of probe extension, levels of background intensity and overall performance of the assay as discussed in Chapter 2 section 2.12. As visualised in Figure 5-3, bisulfite conversion rates were high (≥88%).

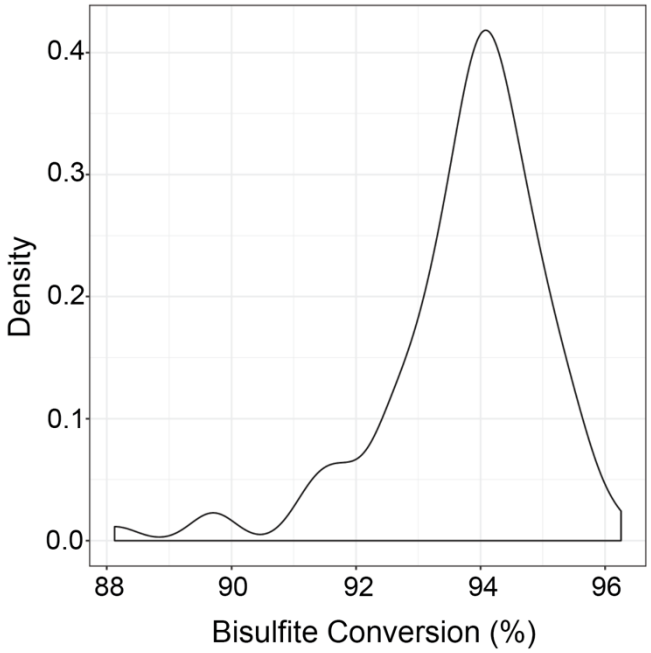


Figure 5-3. Bisulfite conversion efficiency for blood and sperm DNA samples included in the discovery cohort
Conversion efficiences are based on data from the bisulfite conversion control probes included in the Illumina MethylationEPIC array.

5.4.10 Discovery Cohort Quality Control: Outlier Analysis

An outlier analysis was performed using the outlyx() function in wateRmelon based on 1) the interquartile range of the first two principal components 2) the pcoout() algorithm detecting outliers in high dimensional datasets (369). Two blood samples in the discovery cohort showed substantial differences in DNA methylation when compared to the blood samples overall. Further investigation into the two particular samples, including analysis using the SNP genotyping probes described in section 5.4.11, revealed that one had been mislabelled, and was reassigned as a sperm sample. Thus, only one participant (sperm and blood) from the original cohort of 48 was removed from further downstream analyses. The outlier analysis following reassigning of the mislabelled sample is shown in Figure 5-4.

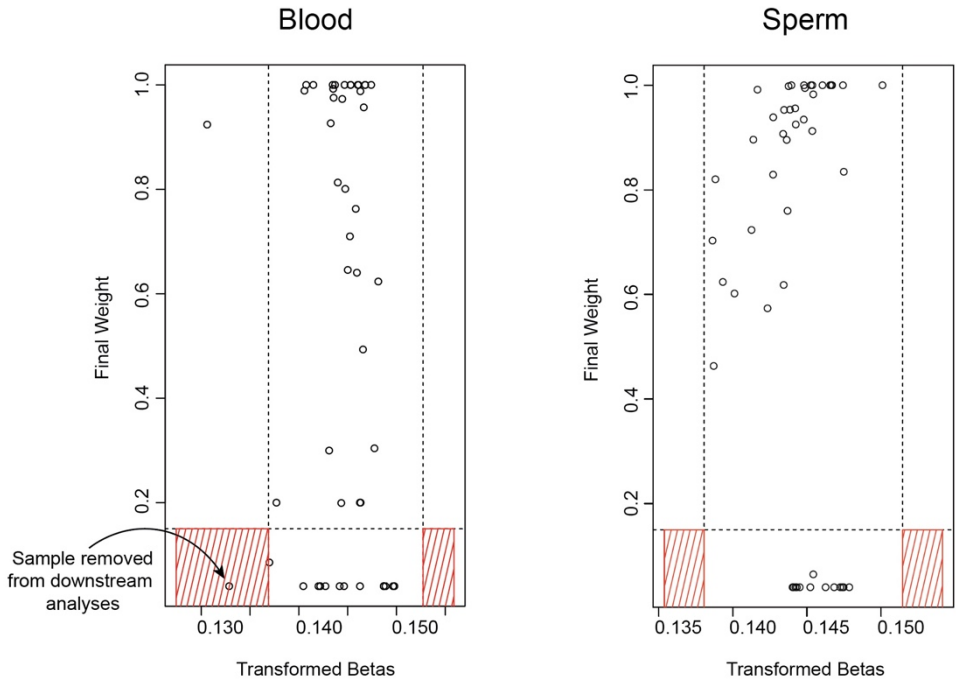


Figure 5-4. Outlier analysis of blood and sperm samples in the discovery cohort.
The sample that appeared as an outlier in two dimensions (as indicated by falling into the area with red, diagonal lines) was removed from further downstream analyses, along with the sperm sample collected from the same individual.

5.4.11 Discovery Cohort Quality Control: SNP based clustering

59 of the control probes included in the Illumina MethylationEPIC array are for direct interrogation of genotype. In this study, which included two matched samples from each individual, information from the genotype probes was used to confirm that the genotypes at these 59 probes are identical for the matched samples. As visualised in Figure 5-5, the clustering dendrogram clustered the samples into pairs and it was confirmed that the two samples from each individual were always clustered into the same pair.

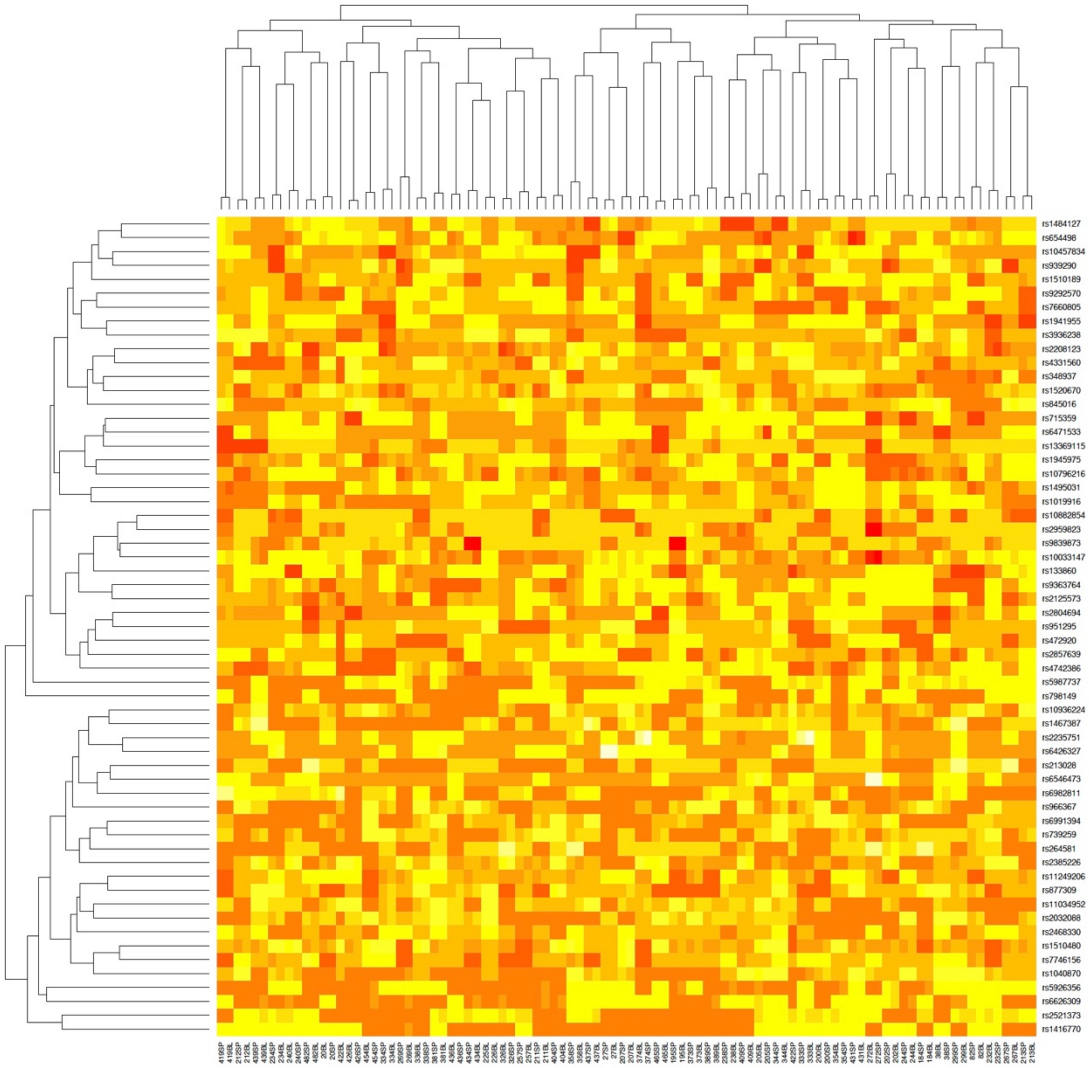


Figure 5-5. Heatmap showing correct clustering of matched samples according to the SNP genotyping control probes.
Matched samples of sperm and blood collected from the same participant show identical genotypes at the 59 genotyping control probes included in the Illumina MethylationEPIC Array.

5.4.12 Discovery Cohort Quality Control: Principal Component Analysis

Principal component analysis identified that the main determinant of differences in DNA methylation in the samples included in the discovery cohort was their tissue of origin. This was expected due to the highly tissue-specific nature of DNA methylation signatures. As visualised in Figure 5-6, blood and semen samples formed two distinct clusters indicating large scale overall differences in DNA methylation.

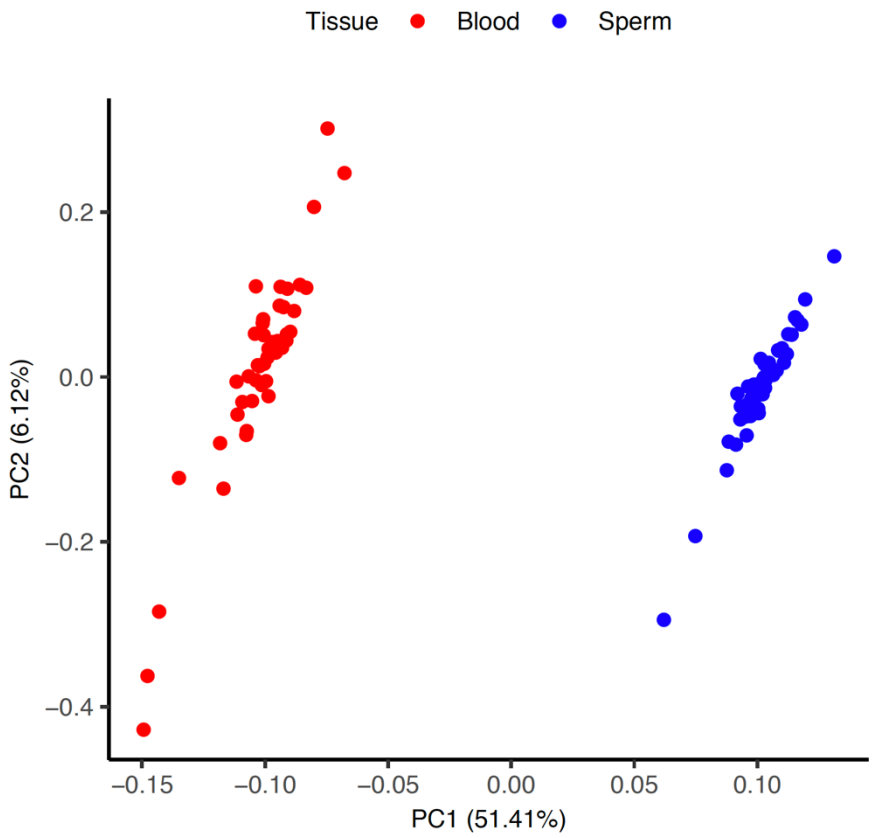


Figure 5-6. PCA plot of samples included in the discovery cohort
As visualised, the tissue of origin accounts for the main variability in the data, which was expected due to the tissue-specific nature of DNA methylation signatures. PCA= Principal Component Analysis, PC1 = Principal Component 1, PC2 = Principal Component 2.

5.4.13 Discovery Cohort Data Preprocessing: Probe Filtering

General considerations behind probe filtering is discussed in Chapter 2 section 2.12.

Prior to probe filtering, a total number of 866,577 probes were included in the MethylationEPIC array analysis. In the case of the discovery cohort, 9779 probes were removed because more than 5% samples displayed a detection p value (probability that the total intensity for a given probe falls within the background signal intensity) higher than 0.05. 3337 probes were removed because of having a bead count of less than three. There were no samples for which more than 5% of CpG sites displayed a detection p value higher than 0.05. Following removal of the probes described above, a total of 853,461 probes were included in further analyses.

The presence of single nucleotide polymorphisms (SNPs) in close proximity to interrogated CpG sites can lead to false assumptions about the association of methylation of a particular CpG site and the outcome variable. In addition, potentially cross-hybridising probes that bind non-specifically to the target regions of interest or that bind to repetitive regions of the genome should be filtered out. Stringent filtering of probes in close proximity (within 10 base pairs) of SNPs as well as potentially cross-reactive probes were filtered using three sets of annotated lists (87, 370, 371). This led to the removal of a total of 149,105 CpG sites. Thus, following probe filtering based on p-values, SNPs in the probe sequences and cross-hybridisation, a total of 704,356 CpG sites were included in further analyses.

5.4.14 Discovery Cohort Data Preprocessing: Data Normalisation

In order to remove technical and systematic variability and to make results comparable across samples, the DNA methylation data was normalised using the dasen method (see Chapter 2 section 2.12 for details). Dasen is a quantile normalisation algorithm which normalises type I and type II backgrounds separately in a first step and then quantile normalises methylated and unmethylated signal intensities. This normalisation also accounts for the different types of assay, the Type I and Type II assays (described in section 2.12) employed in the Illumina MethylationEPIC Array. An example of the density distribution of the red and green

colour channels before and after normalisation using the dasen method is provided in Figure 5-7.

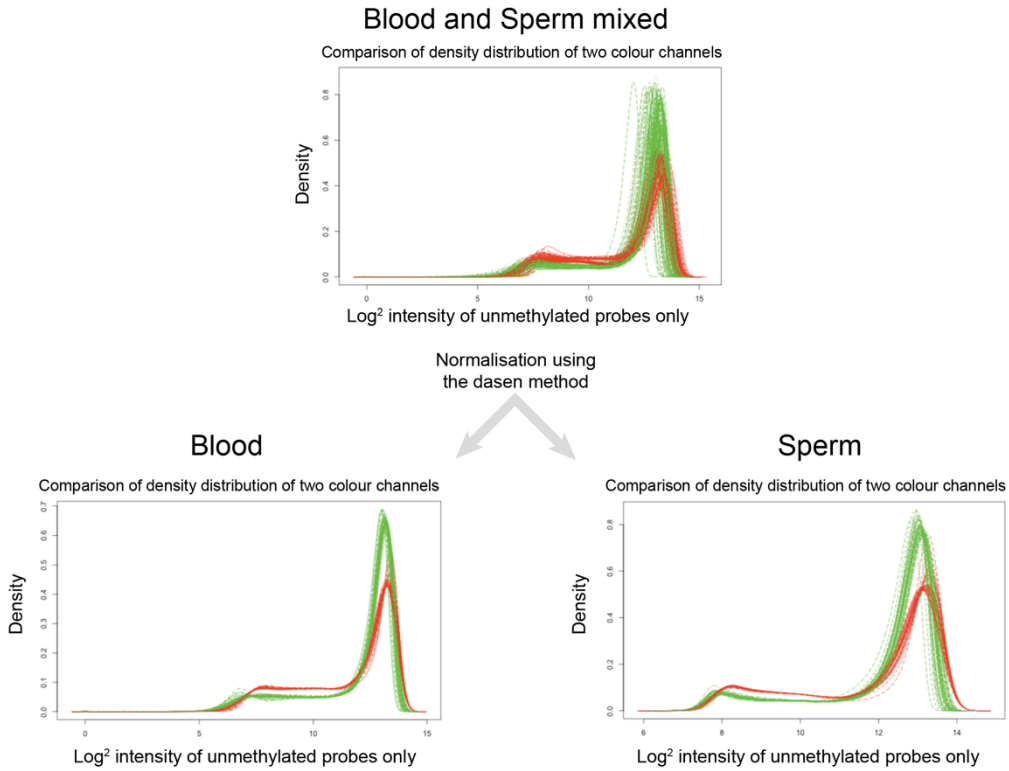


Figure 5-7. Example of pre- and post- dasen normalisation profiles of unmethylated probes from sperm and blood sample methylation data in the discovery cohort.
The red and the green represent the two channels of differing wavelength used to scan the Illumina MethylationEPIC array BeadChips. The x axis depicts Log^2 transformed β values of unmethylated probes and the y axis depicts density. As visualised, data was normalised separately for the two tissues.

5.4.15 Discovery Cohort Data Preprocessing: Methylation Age Prediction

The DNA Methylation Age prediction tool developed by Horvath et al uses DNA methylation data from analyses of close to 8,000 tissue samples to predict the age of study participants based on DNA methylation markers in their tissue samples (196). As visualised in Figure 5-8, the DNA Methylation Age tool accurately predicted the age of study participants when examining their blood samples ($R = 0.74$, $p = <0.001$), but failed to do so in the case of sperm samples ($R = 0.26$, $p = 0.07$).

It should be said that Horvath specifically mentions that the tool inaccurately predicts age in the case of sperm, which is consistent with our findings (196). Therefore, in

subsequent age related analyses of sperm samples (section 5.9), the more recently developed DNA methylation clock developed by Jenkins et al and which was specifically trained on sperm samples, was used in preference to the Horvath model (372).

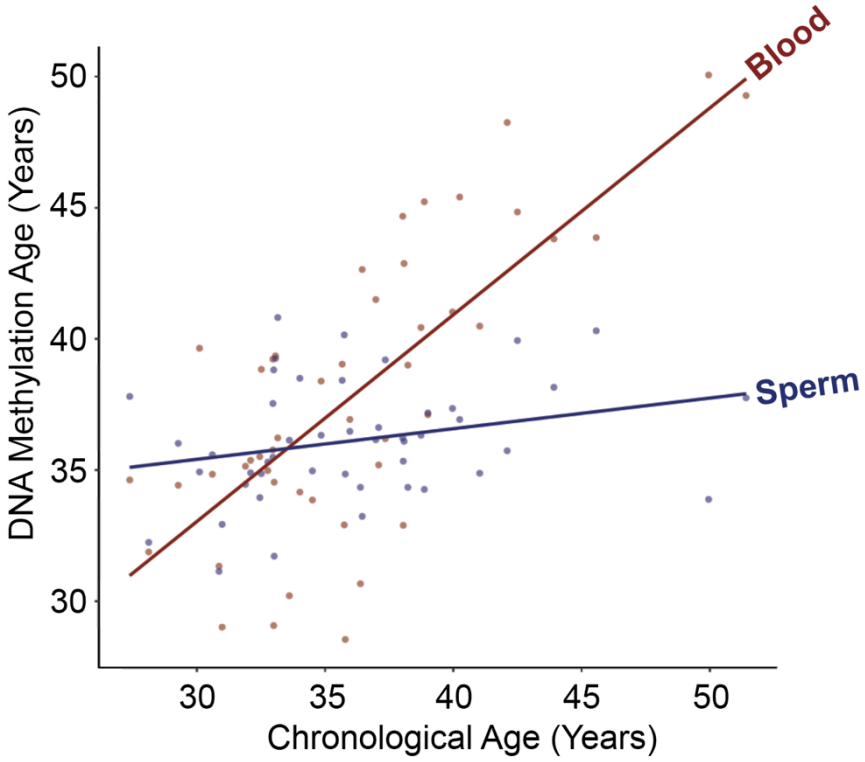


Figure 5-8. Quality control of sperm and blood samples included in the discovery cohort using the DNA Methylation Age prediction tool.
The x axis depicts the age of the participants included in the cohort and the y axis describes their predicted age based on the DNA Methylation Age prediction tool. There was a significant correlation between actual and predicted age of participants when analysing the blood samples (R = 0.74, p = <0.001), however this was not the case for the sperm samples (R = 0.26, p = 0.07).

5.4.16 Summary of Quality Control and Data Preprocessing for the Discovery Cohort

Quality control of DNA methylation data demonstrated a high bisulfite conversion efficiency and that the major explanation for variance between samples was their tissue of origin. One matched set of blood and sperm was removed from further analysis as it was an outlier. Stringent p value filtering and filtering for cross-reactive and SNP probes was applied to minimise misinterpretation of DNA methylation data

in downstream analyses. Use of the DNA Methylation Age prediction tool accurately predicted the age of participants in the case of blood samples, but not in the case of sperm samples.

5.5 Data Preprocessing and Quality Control for the Replication and Obesity Cohorts

Quality control and data preprocessing for the replication and obesity cohorts was performed in the same way as described in for the discovery cohort. The replication and obesity cohorts were analysed on the same DNA MethylationEPIC array such that the array contained 24 matched sets of blood and sperm from overweight/obese males and 24 matched sets of blood and sperm from lean males. They were therefore also analytically processed together. The samples were randomised on the array to minimise any potential effect of sample location on array processing and results.

5.5.1 Replication and Obesity Cohorts Quality Control: Bisulfite Conversion Efficiency and Evaluation of Control Probes

An initial quality control report from UCL Genomics, where the MethylationEPIC arrays were processed, reported high performance of the sample dependent and sample independent quality control probes (data not shown).

The sample dependent controls (bisulfite conversion control probes and probes that assess the specificity of probe extension, levels of background intensity and overall performance of the assay also demonstrated good performance. Bisulfite conversion rates were high (≥85%).

5.5.2 Replication and Obesity Cohorts Quality Control: Outlier Analysis

An outlier analysis of samples included in the replication and obesity cohorts detected no outliers and thus no samples were excluded from further analyses (Figure 5-9).

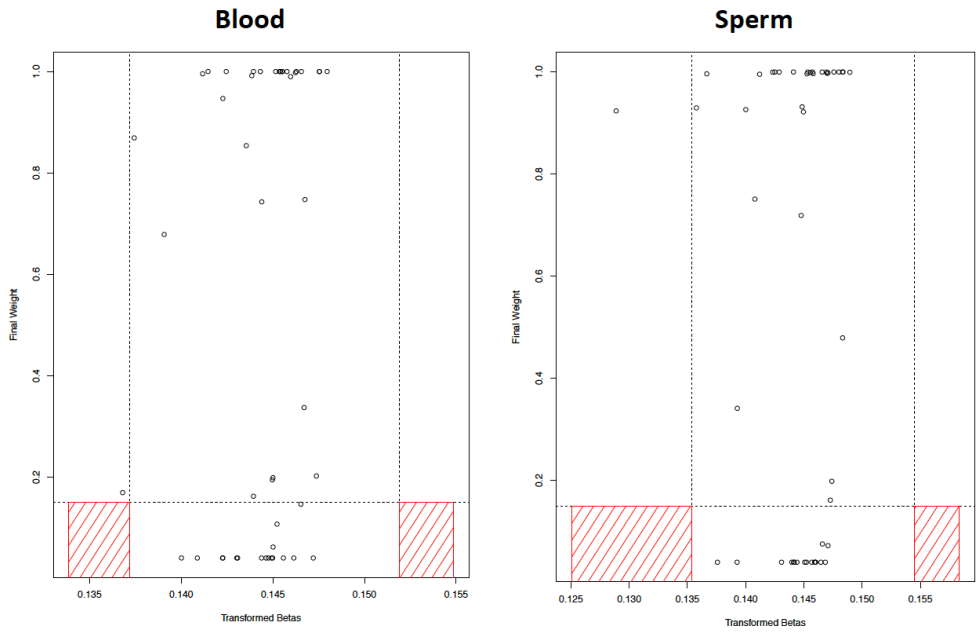


Figure 5-9. Outlier analysis of blood and sperm samples included in the replication and obesity cohorts.

5.5.3 Replication and Obesity Cohorts Quality Control: SNP based clustering

SNP based clustering based on the 59 SNP probes included in the DNA Methylation EPIC array identified five problematic samples, three of which were samples where blood DNA matched semen of another study participant, and two of which were samples that did not match any other samples in the study. This suggested mix-ups in the collection or laboratory handling procedures and these samples were therefore removed from downstream analyses. The problematic samples were from two participants in the replication cohort and three participants in the obesity cohort. Thus, the replication and obesity cohort study populations were reduced to 21 and 22 participants respectively.

5.5.4 Replication and Obesity Cohorts Data Preprocessing: Probe Filtering

In the case of the array on which the replication and obesity cohorts were processed, 11,366 CpG sites were removed from downstream analysis due to having a bead count of less than three. 9,443 CpG sites were removed due to having a detection p-value > 0.05. A further 148,400 CpG sites were removed due to being in close proximity (within 10 base pairs) of SNPs, or for being potentially cross-hybridising

probes (87, 370, 371). This left a total number of 697,442 CpG sites that were included in downstream analyses.

5.5.5 Replication and Obesity Cohorts Data Preprocessing: Data Normalisation

As for the discovery cohort, data was normalised using the dasen method.

5.5.6 Replication and Obesity Cohorts Data Preprocessing: Methylation Age Prediction

As was the case for the discovery cohort (section 5.4.15), the DNA Methylation Age prediction tool accurately predicted the chronological age of participants in the case of blood samples ($R = 0.82$ and $p = 2.355 \times 10^{-11}$), but not in the case of sperm samples ($R = -0.01$ and $p = 0.967$) in both the replication and obesity cohorts (196).

5.5.7 Summary of Quality Control and Data Preprocessing for the Replication and Obesity Cohorts

Quality control of DNA methylation data generated by DNA MethylationEPIC array analysis of 24 matched samples of blood and sperm from lean males and 24 matched samples of blood and sperm from overweight/obese males demonstrated a high bisulfite conversion efficiency and that the major explanation for variance between samples was their tissue of origin. Five matched samples of blood and sperm were removed from further downstream analysis due to incorrect SNP-based clustering. Stringent p value filtering and filtering for cross-reactive and SNP probes was applied to minimise misinterpretation of DNA methylation data in downstream analyses. Use of the DNA Methylation Age prediction tool accurately predicted the age of participants in the case of blood samples, but not in the case of sperm samples.

5.6 DNA Methylation Analysis Results

5.6.1 Accounting for cell type composition in blood samples

Buffy coat derived from whole blood represents a heterogenous tissue containing numerous different types of leukocytes. This mixed cell type composition can induce confounding in epigenetic analyses if not accounted for. Thus, blood cell type coefficients of monocytes, granulocytes, NK-cells, B cells, CD8+-T-cells, and CD4+-T-cells were estimated from the methylation data using the method described by Houseman et al (214) (Figure 5-10)

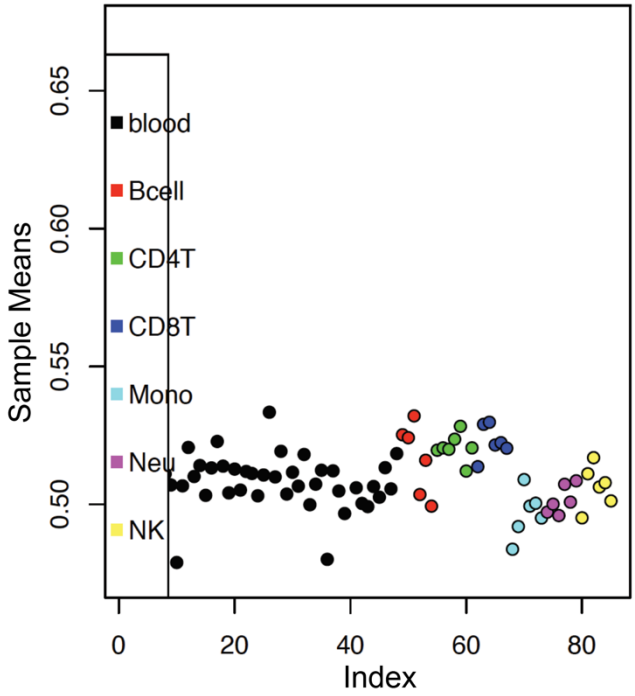


Figure 5-10. Blood cell DNA methylation coefficients for blood samples included in the discovery cohort
The black dots represent the average DNA methylation level at cell-type discriminating probes in the discovery cohort samples. The coloured dots represent the average DNA methylation levels in the reference sorted cells that were used to determine the prediction model
CD4T = CD4 T-cell, CD8T = CD8 T-cell, Neu = Neutrophils, NK = Natural Killer cells.

5.6.2 General characterisation of sperm and blood methylomes

Comparison of the array-wide distribution of CpG methylation in sperm versus blood revealed that sperm exhibits a highly polarised methylation profile towards the two extremes of DNA methylation levels (Figure 5-11). Thus, both low (<20%) and high (>80%) levels of methylation were more commonly seen in sperm than in blood, which displayed a tendency towards more intermediate levels of methylation (20-80%). Overall, the sperm methylome was found to be slightly hypermethylated compared to blood (independent sample t-test on median DNA methylation values in blood and sperm, P = 0.016, mean difference in DNA methylation = 0.14%). Overall levels of methylation showed substantial differences between sperm and blood. 603,519 probes were significantly differentially methylated between the two tissues (paired t-test, FDR < 0.05). At 349,951 (58%) of these significant sites, hypermethylation of sperm compared to blood was observed, while 252,242 (42%) showed lower methylation in sperm compared to blood.

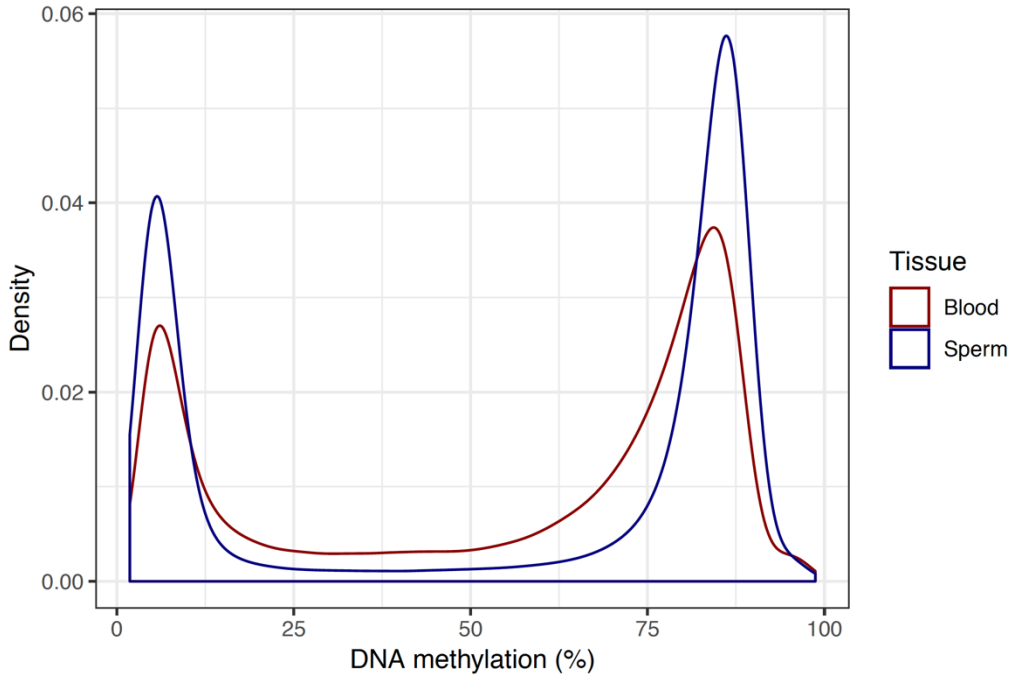


Figure 5-11. Genome-wide comparison of CpG methylation in sperm and blood for the discovery dataset
Comparison of genome-wide CpG methylation between blood and sperm revealed that the sperm methylome is highly polarised, such that both low (<20%) and high (>80%) methylation levels are more common in sperm when compared to blood.

5.6.3 Proportion of DNA methylation variance explained

The extent to which blood can be used as a proxy tissue for sperm DNA methylation was then addressed by assessing the proportion of DNA methylation variance in sperm that can be explained by the DNA methylation variance in blood. This relationship is presented in Figure 5-12. As shown, for the majority of CpG sites, interindividual methylation variation in blood explains only a small amount of the variation seen in spermatozoa.

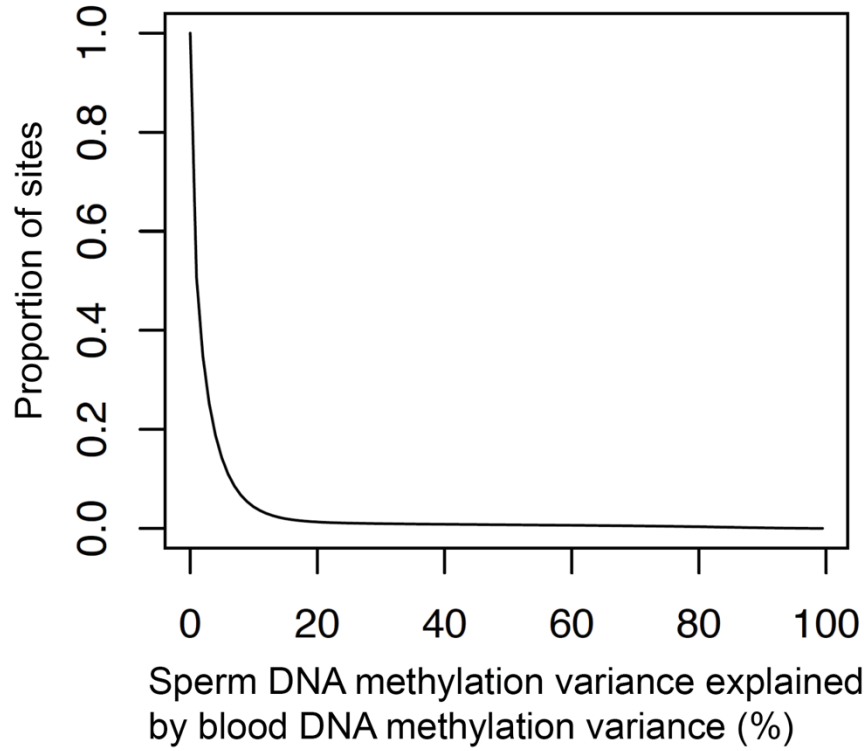


Figure 5-12. Variation in DNA methylation in blood as a predictor of DNA methylation variation in sperm
Shown is the proportion of sites (y-axis) for which variation in blood explains a certain of percentage of DNA methylation variance (x-axis) in sperm

A linear regression model was used to calculate the proportion of variance in DNA methylation explained by tissue (sperm versus blood), individual and age. As shown in Figure 5-13, tissue was able to explain the largest proportion of variance in DNA methylation compared to age and individual.

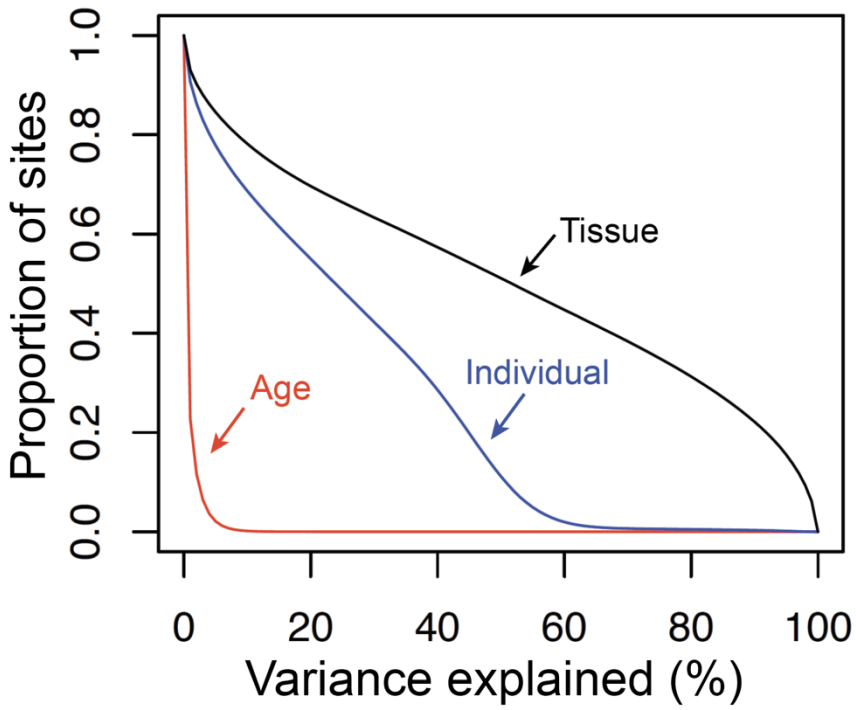


Figure 5-13. Proportion of DNA methylation variance in whole blood and sperm that is explained by tissue, individual and age

5.6.4 Characterisation of sperm and blood methylomes according to genomic regions

A more detailed characterisation of the relative levels of CpG methylation across different genomic regions was then performed. In this, DNA methylation levels were assessed in CpG islands (500-1500 base pair long sequences with a an observed-to-expected CpG ratio greater than 0.6), CpG island shores (regions with lower CpG density that lie within 2 kb up- and downstream of a CpG island), CpG shelves (regions located 2 kb outside of a CpG island shore) and CpG sites in open seas (CpGs not associated with an CpG island) (373). To compare DNA methylation levels between sperm and blood in these genomic regions, the mean methylation at each CpG site was calculated and a two sample t-test was used to compare the means between the tissues. In other words, the mean of the mean methylation level at all CpG sites assigned to CpG islands in sperm was compared to the mean of the mean methylation level at all CpG sites assigned to CpG islands in blood and so forth.

As shown in Figure 5-14, there were clear differences in CpG methylation levels between blood and sperm in CpG islands, CpG island shores, CpG island shelves and CpG sites in open sea in blood and sperm ($p < 2.2 * 10^{-16}$ for all of these). In CpG islands, CpG methylation levels in blood were ~7% higher than in sperm (132,883 probes assessed). In CpG island shores, CpG methylation levels in blood were ~16% higher than in sperm (128,079 probes assessed). In CpG island shelves, CpG methylation levels in blood were ~6% lower than in sperm (48,301 probes assessed). At CpG sites in open seas, CpG methylation levels in blood were ~7% lower than in sperm (395,093 probes assessed).

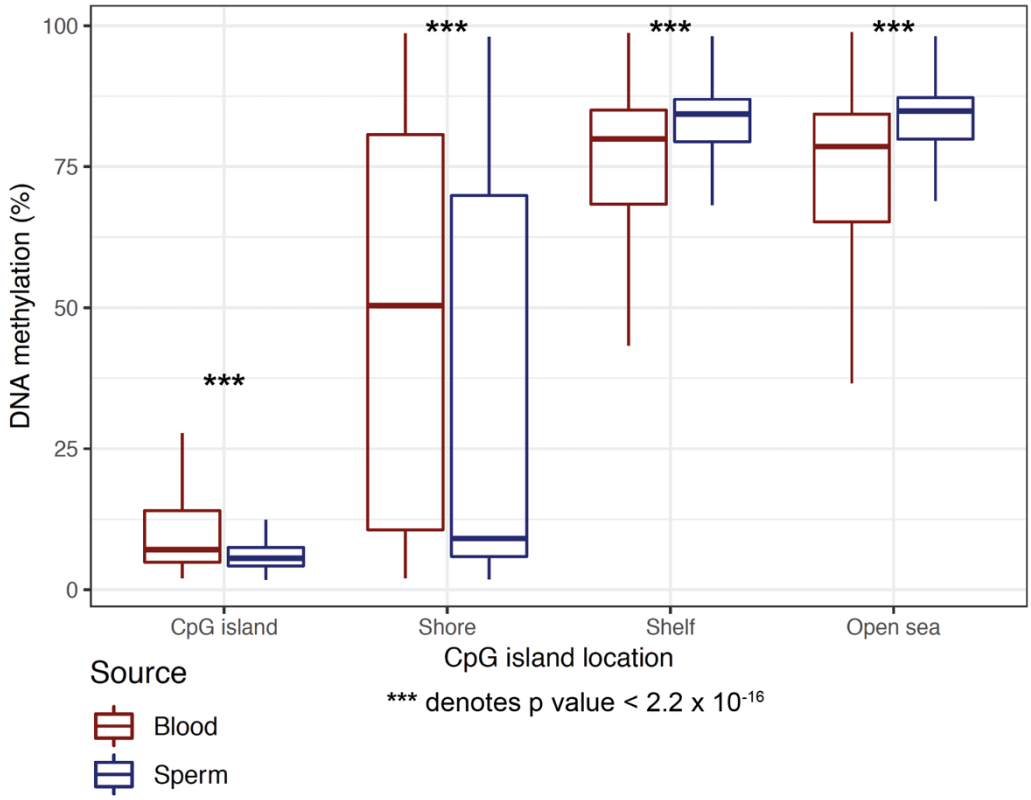


Figure 5-14. Comparison of DNA methylation levels in CpG islands, CpG island shores, CpG island shelves and CpG sites in open seas in blood and sperm

A comparison of CpG methylation levels between blood and sperm in regions 200 base pairs from transcription start sites (TSS200), regions 1500 base pairs from transcription start sites (TSS1500), 5' untranslated regions (5' UTR), gene bodies, 3' untranslated regions (3' UTR) and CpG sites that are not annotated to genomic regions is presented in Figure 5-15. In TSS200 and TSS1500 regions, the methylation levels in blood were significantly higher than those in sperm ($p < 2.2 * 10^{-16}$ for both). CpG methylation in TSS200 regions was ~2% higher in blood (54,041 probes assessed) and CpG methylation in TSS1500 regions was ~11% higher in blood (88,587 probes assessed). CpG methylation in 5' UTRs was lower in blood by ~2% ($p < 5.3 * 10^{-14}$; 61,449 probes assessed). CpG methylation in gene bodies was also lower in blood by ~2% ($p < 2.2 * 10^{-16}$; 290,256 probes assessed). CpG methylation in 3' UTRs was not significantly different between blood and sperm ($p = 0.830$; 17,923 probes assessed). In CpG sites not annotated to genomic regions, CpG methylation was lower in blood by ~2% ($p < 2.2 * 10^{-16}$; 192,100 probes assessed).

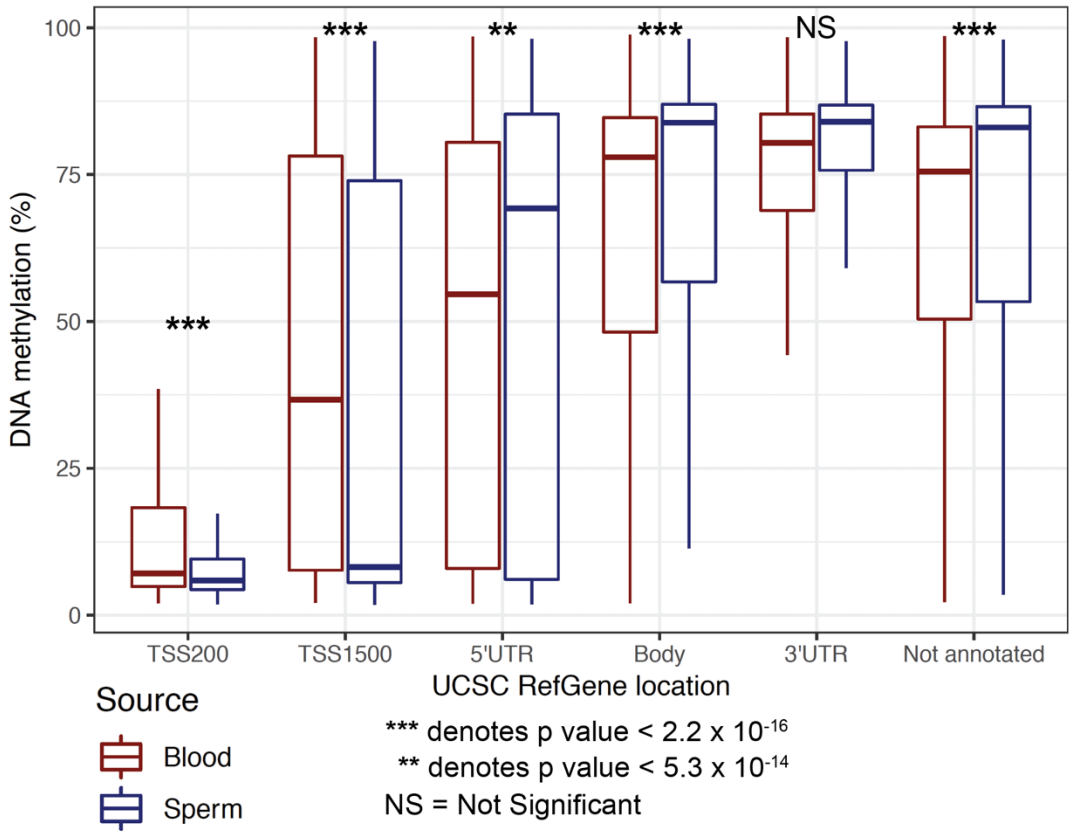


Figure 5-15. Comparison of DNA methylation levels in blood and sperm in regions 200 base pairs from transcription start sites (TSS200), regions 1500 base pairs from transcription start sites (TSS1500), 5' untranslated regions (5' UTR), gene bodies, 3' untranslated regions (3' UTR) and CpG sites that are not annotated to genomic regions

5.7 Detailed Analysis of the Correlation of DNA Methylation between Blood and Sperm

The correlation of DNA methylation between blood and sperm was first explored by comparing the observed distribution of correlation coefficients against a simulated null distribution. In this, samples were randomly permuted and correlations between DNA methylation in whole blood and sperm were recalculated. Had there been no methylation correlation between the two tissues, the observed distribution of correlation coefficients would align to the null distribution. As presented in Figure 5-16, the observed distribution of correlation coefficients closely matched that of the null distribution. However, the histogram was shifted slightly to the right, i.e. there was a slightly higher correlation between the tissues than what would have been expected if the tissues were entirely uncorrelated.

Interestingly, there was also a small number of specific CpG sites that appeared to display a high correlation of DNA methylation levels between the two tissues (circled in green in Figure 5-16). The number of CpG sites where the DNA methylation level was significantly correlated between sperm and blood was calculated using the empirically derived significance threshold for MethylationEPIC array analyses of $9 * 10^{-8}$ (366). Using this threshold, a total of 5,307 CpG sites (i.e. approximately 0.8% of the 704,356 informative sites in the discovery cohort) were found to have significantly correlated methylation levels between the tissues. This number was similar to if a Bonferroni adjusted significance had been used (5,272 sites) but slightly lower than if an FDR adjusted significance threshold had been used (7,856 sites).

Of these 5,307 significantly correlated CpG sites, 5,271 were also represented in the replication/obesity quality controlled dataset. The 5,307 sites were taken forward for further characterisation, partly because these could potentially be sites were using blood as a proxy tissue for sperm DNA methylation could be justified.

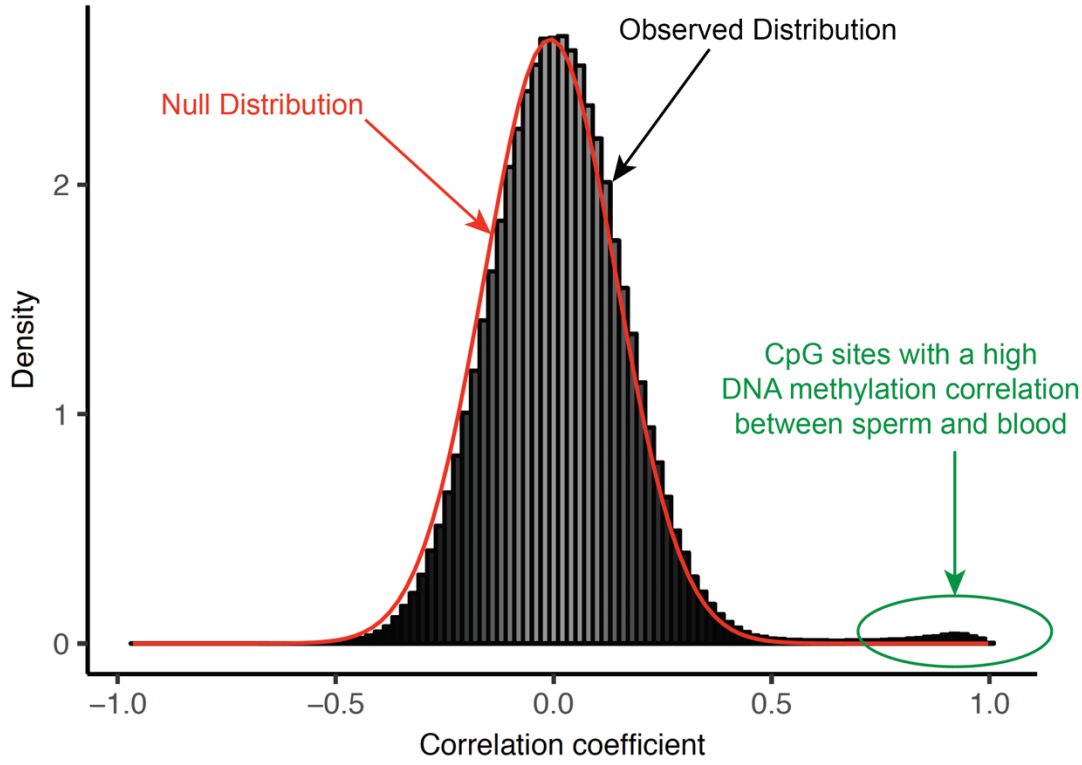


Figure 5-16. Histogram showing the distribution of correlation coefficients between DNA methylation in whole blood and sperm (discovery dataset only)

5.7.1 Characterisation of CpG sites with apparently high methylation correlation between sperm and blood

The 5,307 CpG sites that appeared to display a high level of methylation correlation between the two tissues were characterised first by visually inspecting the distribution of DNA methylation levels. From doing so, it appeared that although stringent SNP filtering criteria had been applied (section 5.4.13), methylation levels in a large proportion of these CpG sites appeared to be genetically driven. This was suggested from the way the majority of these CpG sites displayed bimodal and trimodal patterns of methylation, as exemplified in Figure 5-17.

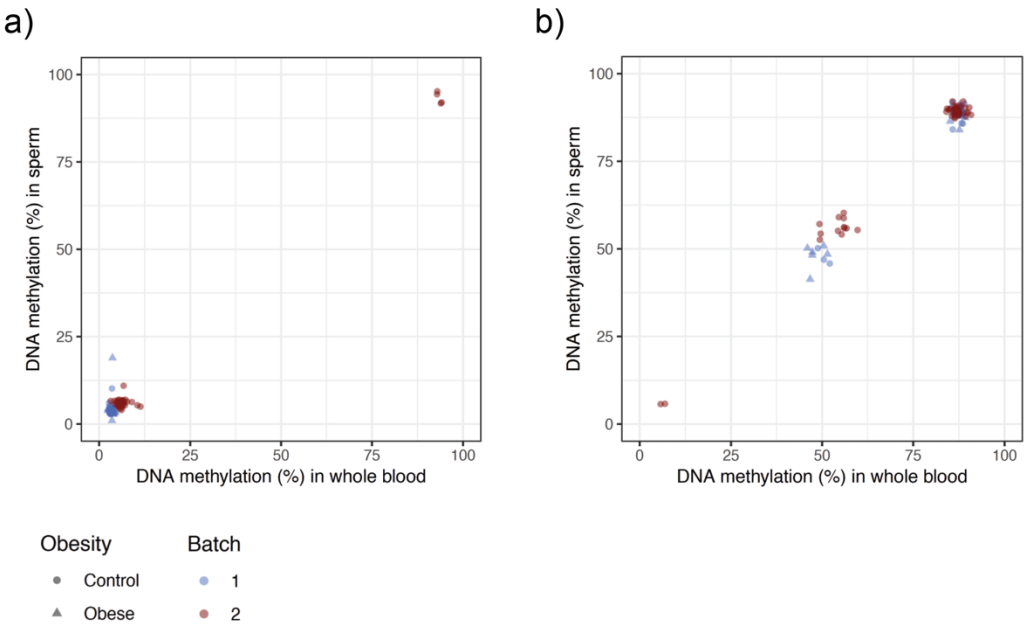


Figure 5-17. Examples of bimodal and trimodal DNA methylation distributions when comparing blood and sperm
a) cg24596064, which displayed a bimodal pattern of methylation level distributions
b) cg14499602, which displayed a trimodal pattern of methylation level distributions

The presence of SNPs in the CpG site and probe sequence is a recognised complication in interpreting DNA methylation data. To further explore this issue in the analysis of the 5,307 sites that appeared to display a high methylation correlation between blood and sperm, two approaches were applied. First, an algorithm was devised such that CpG sites where methylation levels formed two or three distinct clusters (bimodal/ trimodal distribution patterns) were identified. In this, a two

dimensional outlier test was used by adapting the rosnerTest() function in R to exclude unimodal distributions. Next, k means clustering was applied for 2 and 3 clusters as implemented in the function pamk() of the R package *cluster*. This function determines the best fitting number of clusters (two or three – corresponding to bi- and tri-modal methylation distributions). The clustering algorithm identified 5052 with a bimodal pattern of methylation, and 255 sites that displayed a trimodal pattern of methylation distribution, i.e. there were no CpG sites with a unimodal methylation distribution. Second, the CpG sites that correlated between tissues were cross-checked for SNPs in the CpG site using the dbSNP Human Build 151 database (374). Only 11 out of the 5,307 significantly correlated sites were found to have no known SNPs anywhere in the probe sequence. Of these, 10 were represented among the quality controlled probes in the obesity/replication cohorts.

The CpG sites where two clusters were identified by the clustering algorithm were explored further by assessing the number of individuals in the smallest cluster. In 3078 out of the 5052 sites (~61%) with two clusters, the smallest cluster consisted of only one individual outlier. This was also the case when examining CpG sites that displayed a trimodal distribution of DNA methylation levels. Such findings are consistent with these individual outliers having a rare or low frequency genetic variant that influences DNA methylation levels.

It was then investigated whether the same or only a small number of individuals were overrepresented in the clusters that consisted of only one or a couple of individuals. This analysis also included data on the self-reported ethnicity of participants, as a probable cause for genetic variation would be different ethnic background. However, as shown in Figure 5-18, although a couple of individuals had either no instances of being the sole outlier in a bi- or trimodal pattern of methylation distribution and a couple of individuals were very frequently outliers, it was more commonly the case that different individuals were outliers at different CpG sites. Further, there was no correlation between the frequency of private clusters and the individual's reported ethnicity (data not shown).

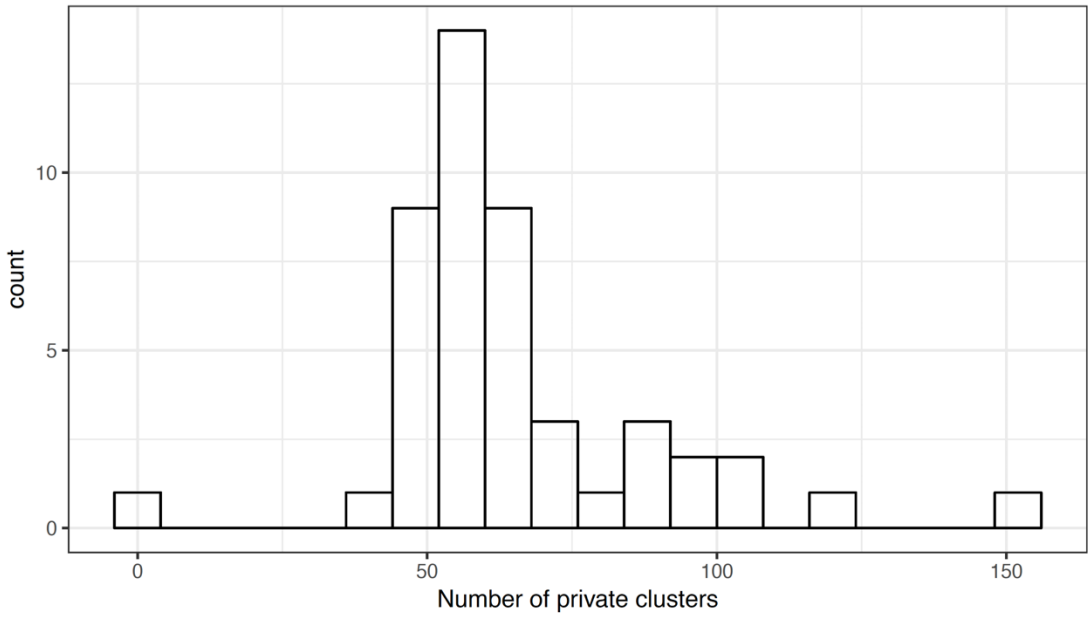


Figure 5-18. Analysis of whether some individuals were overrepresented in private clusters

In summary, it seemed that for the vast majority of CpG sites that originally appeared to exhibit DNA methylation co-variation between sperm and blood, the correlation was likely driven by the underlying genotype. Examples of methylation data for the 10 CpG sites where no known SNP in the CpG site was able to explain the high DNA methylation correlation between the tissues (in the discovery and replication/obesity cohorts) are presented in Figure 5-19. For some of these sites, such as cg21625068 in Figure 5-19 a), there did indeed appear to be a true correlation between sperm and blood DNA methylation levels in a non-genetic manner. For other sites, however, such as cg23350716 in Figure 5-19 b), the methylation levels seemed to still cluster in a pattern consistent with underlying genetic effects. It is possible that this effect could have been caused by SNPs in the probe sequence not present on the comprehensive dbSNP Human Build 151 database, or represent a true strong biological effect of a SNP not in the probe sequence at the CpG site of interest (374).

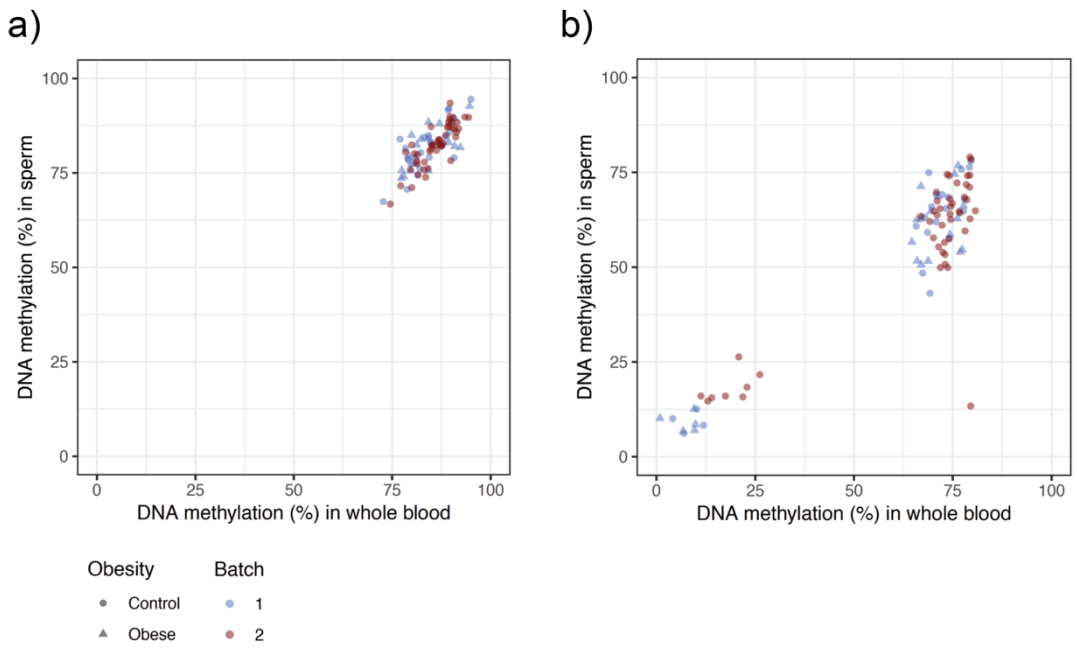


Figure 5-19. Examples of CpG sites where the methylation correlation between sperm and blood could not be explained by known SNPs in the CpG site.
These scatter plots depict all samples included in the study, i.e. from the discovery, replication and obesity cohorts. 'Batch' refers to whether the samples were from the discovery cohort ('Batch 1') or the replication and obesity cohorts ('Batch 2'). Whether the participant was lean or obese is indicated by the shape of the marker.
a) cg21625068
b) cg23350716

5.7.2 Sites characterised by a negative DNA methylation correlation between sperm and blood

A subset of CpG sites among the 5,307 that were significantly correlated in the discovery cohort (n = 44 at the empirically derived threshold for MethylationEPIC arrays of 9 * 10^{-8}, 42 of which were included in the quality controlled replication/obesity dataset) displayed negative correlation between sperm and blood. In other words, for these sites, the higher the DNA methylation levels were in blood, the lower the methylation levels were in sperm. Shown in Figure 5-20 are two examples of such negative DNA methylation correlation; cg05480191 (Figure 5-20 a) and cg01812571 (Figure 5-20 b).

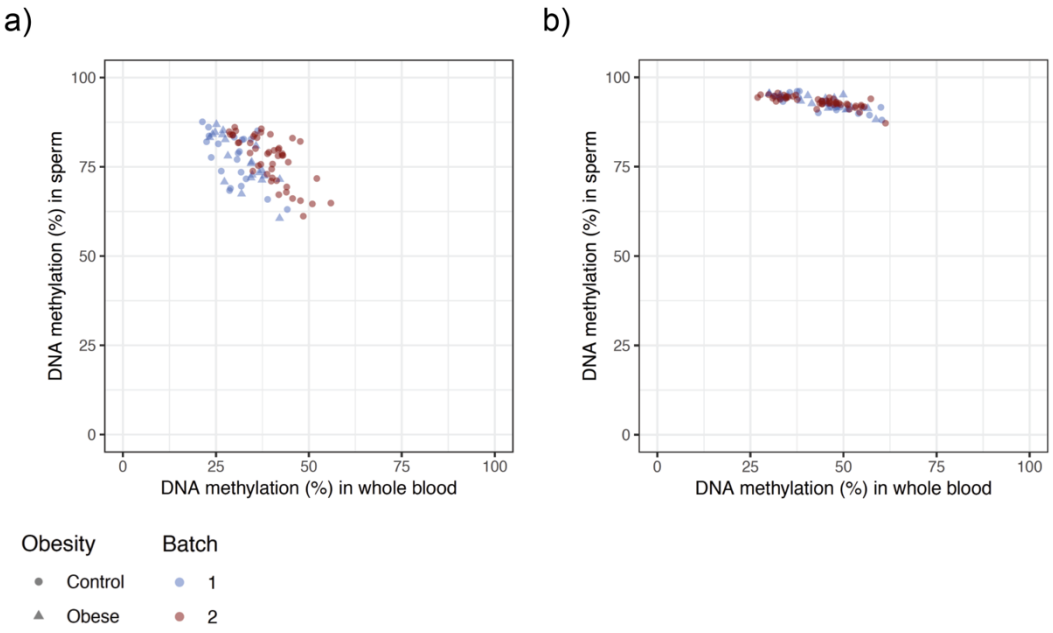


Figure 5-20. Examples of sites of significant negative DNA methylation between blood and sperm.
a) cg05480191
b) cg01812571

5.7.3 Correlated sites characterised by highly discordant methylation between sperm and blood

A high proportion of CpG sites among the 5,307 in the discovery cohort that were significantly correlated showed both significant correlation and significant differences between blood and sperm DNA methylation (n = 3,348). This was calculated using a paired t-test for mean methylation at the 5,307 probes. Two examples of such sites are provided in Figure 5-21.

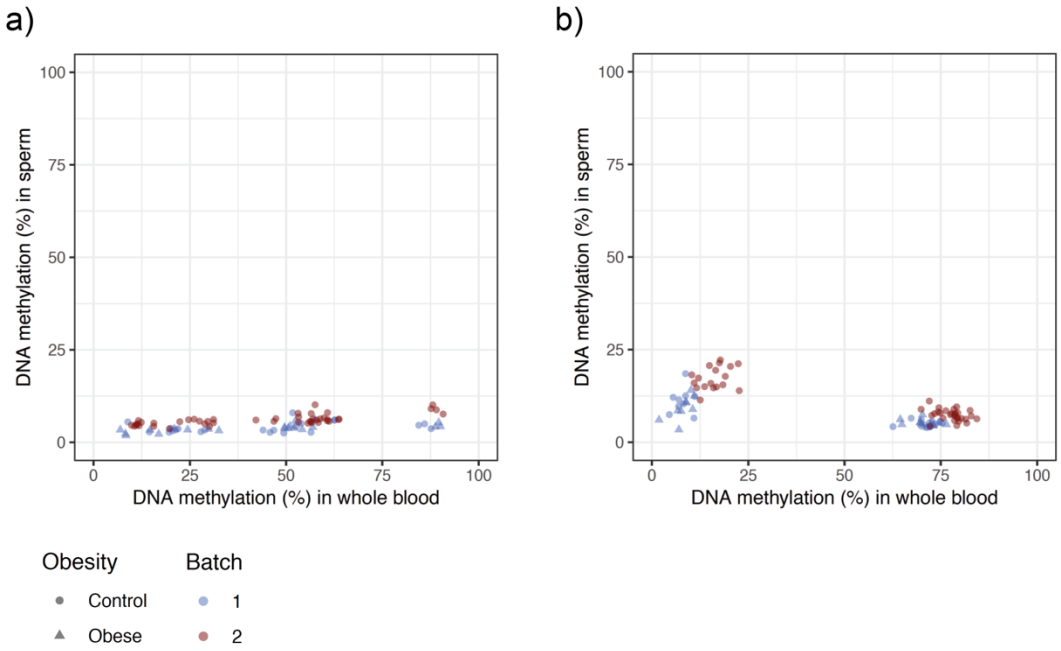


Figure 5-21. Examples of correlated CpG sites that displayed highly discordant DNA methylation between sperm and blood
a) cg07533224, where the methylation level in sperm is consistently low, but the methylation level in blood appears to follow a trimodal distribution pattern
b) cg00456343, where the methylation level in sperm is consistently low, but the methylation level in blood appears to follow a bimodal distribution pattern

In the case of cg07533224 (Figure 5-21 a)), the DNA methylation level in sperm is consistently low (<15%) for all participants, whereas the methylation level in blood appears to follow a trimodal distribution pattern with methylation levels ranging from approximately 5% to approximately 85%. It is possible that tissue specific transcription factors (TFs) are responsible for repressing particular genomic regions via DNA methylation according to tissue specific gene expression requirements. Indeed, it is well established that TFs can influence the establishment and maintenance of DNA methylation at particular genomic regions (375).

5.8 Comparison of results with the obesity cohort

To make use of the full collection of samples and explore the relationship between obesity, genetics and DNA methylation we analysed two further models. These were run on 692,265 probes that remained after filtering in both datasets (discovery and replication/obesity).

First, a mixed effects model was run regressing DNA methylation onto tissue (blood versus sperm), age, batch and obesity status, while controlling for interindividual variation with a random effect:

fit <- lmer(Meth ~ Tissue + Age + Batch + Obesity +(1|ID), data=pheno, REML=F)

As in the discovery cohort, at a large proportion of sites, DNA methylation was associated with tissue type. Using the empirically derived threshold of $9 * 10^{-8}$, 444,785 out of the 692,265 probes (64%) showed significant DNA methylation differences between the blood and sperm. In 274,854 (62%) of these, DNA methylation in sperm was higher than that in blood.

This model identified that methylation at one CpG site, cg19357369, was significantly associated with obesity status in sperm and blood ($p = 8.95 * 10^{-8}$) (Figure 5-22). The effect size was 1.4%, i.e. the average DNA methylation difference between lean and obese participants at this site was 1.4%.

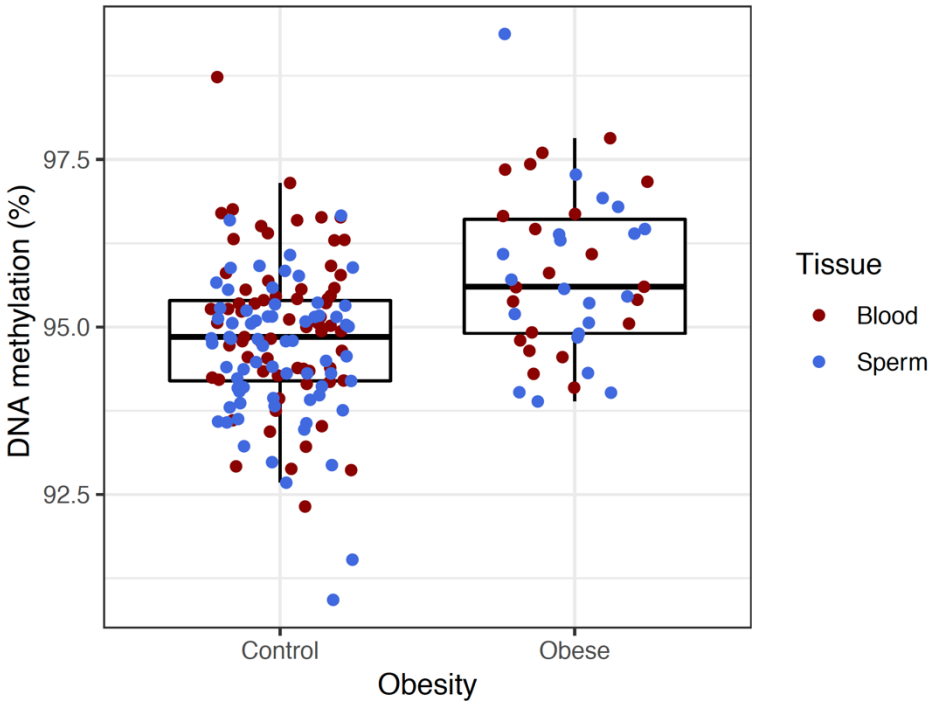


Figure 5-22. Boxplot showing methylation levels in control participants (discovery and replication cohorts combined) compared to obese participants at cg19357369

Secondly, an interaction model was used to determine whether obesity altered the nature of blood and sperm DNA methylation covariation. This model regressed DNA methylation in blood onto DNA methylation in sperm, obesity and their interaction effect, while covarying for experimental batch and age:

$$fit <- lm(Meth_{Blood} \sim Meth_{sperm} * Obesity + Age + Batch)$$

This statistical model identified 691 CpG sites where obesity was significantly associated with blood DNA methylation and 7668 CpG sites where sperm DNA methylation had a significant effect of blood DNA methylation ($p < 9 * 10^{-8}$). Further, there was a significant interaction effect at 686 CpG sites where obesity impacted on the correlation between sperm and blood DNA methylation.

However, there was considerable overlap between the CpG sites identified in this model and those previously identified as significantly correlated between sperm and blood in previous analyses in either the discovery or replication/obesity dataset. Thus, 600 out of 691 CpG sites (87%) where obesity was significantly associated with blood DNA methylation were the same as those previously identified as significantly

correlated between sperm and blood (section 5.7). 6,675 out of the 7,668 CpG sites (87%) where sperm DNA methylation had a significant effect of blood DNA methylation were the same as previously identified as correlated. In the case of the 686 CpG sites where obesity significantly impacted on the correlation between sperm and blood DNA methylation, 596 (87%) were the same as those already identified as correlated. As the clustering algorithm together with the dbSNP data strongly suggested that these correlation effects were mostly driven by genetic variation, this also suggests that the results from the interaction model also are driven by genetic variation.

There was also considerable overlap of sites between the categories of sites identified in the interaction model. Indeed, after removing sites that had previously been identified as significantly correlated between the tissues and those represented in the other categories, there were only 6 sites where obesity significantly impacted on the correlation between sperm and blood DNA methylation. In all of these cases, the effect again seemed to be caused by one or a small number of outliers in a manner suggestive of a genetic effect (exemplified in Figure 5-23).

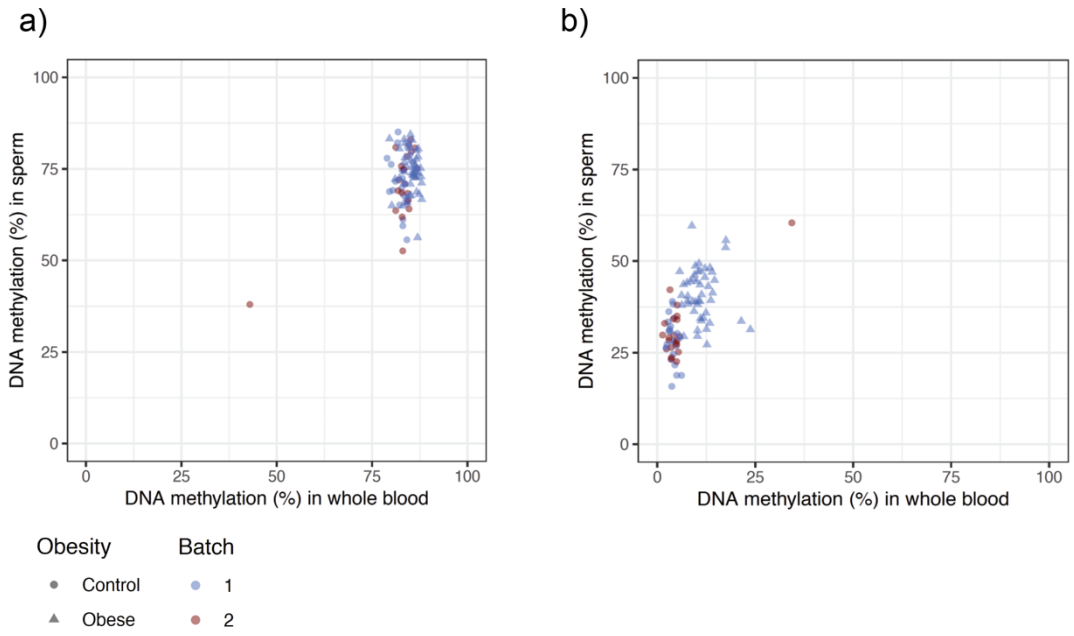


Figure 5-23. Examples of CpG sites where obesity significantly impacted on the correlation between sperm and blood, and which had not been previously identified as correlated between the two tissues
a. cg03263948
b. cg08584107

Similar methylation distributions were observed for the vast majority of CpG sites identified as by the interaction model, i.e. that results were highly suggestive of genetic effects. In conclusion, therefore, there was no convincing evidence that obesity influenced the correlation between sperm and blood.

5.9 Comparison of the sperm methylome to tissues in the Gene Expression Omnibus (GEO) database

The GEO database (Chapter 2 section 2.12.7) was used to identify CpG sites that significantly differ between sperm and a range of human somatic tissues (210). The DNA methylation data was processed using the bigmelon R package and statistical tests were performed using limma (211, 376).

The following criteria were used to select tissue samples on GEO:

- HM450K data available (or Illumina MethylationEPIC data, in which case only the CpG sites represented on the HM450 were included in the model)
- Samples from males only

HM450K methylation data from 281 sperm samples (from five different accessions) was available on GEO. Together with data from the 90 MethylationEPIC samples that passed quality control in the discovery, replication and obesity cohorts, this generated a total sample size of 371 sperm samples. The age of participants was only detailed for six out of the 281 sperm samples available on GEO. Therefore, as the more widely used model by Horvath et al did not provide accurate age estimates for sperm samples (section 5.4.15), a recently developed model by Jenkins et al was used to impute the age of participants from the other 275 samples (196, 372). This model was specifically designed to predict chronological age from genome-wide sperm DNA methylation data. For the sperm samples where age data was actually available (6 from GEO and 90 from the MethylationEPIC dataset), the model accurately reflected the chronological age of participants ($R^2 = 0.13$, $p = 0.0002$) (Figure 5-24).

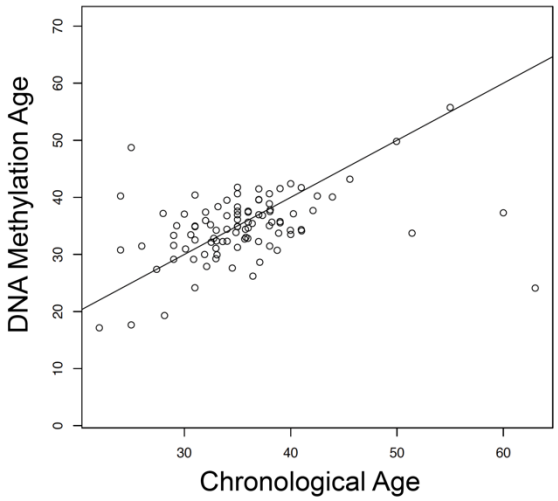


Figure 5-24. Comparison between the DNA methylation age predicted by the sperm-specific DNA methylation age tool developed by Jenkins et al compared to the age of participants where this data was available

DNA methylation data from 5,917 somatic tissue samples which met the above criteria on GEO and were included in the analysis. These are detailed in Table 5-3.

Type of tissue	N	Type of Tissue	N
Adipose	42	Mesenchymal stem cells	9
Amniotic membrane	3	Mesenchymal stromal cells	8
Blood	2,317	Mucosa	95
Brain	978	Muscle	17
Buccal	214	Neuroendocrine tumour	3
Cardiac tissue	11	Neuron	71
Cartilage	60	Neutrophils	69
Chorion	3	Pancreas	112
Chorionic Villi	68	Pituitary adenoma	21
Colon	170	Rectum	13
Cornea	8	Saliva	146
Epithelial	183	Skin	38
Fibroblast	54	Squamous cell carcinoma	7
Glioblastoma	36	T Cells	306
Intestines	1	Thymoma	11
Kidney	45	Thymus	1
Liver	90	Umbilical cord	6
Lung	103	Unsorted Cell Line	9
Lung adenocarcinoma	12	Unsorted Tissues	379
Lymph node	24	Unsorted Tumours	174

Table 5-3. Types of somatic tissue samples included in the GEO analysis

5.9.1 Identification of significant CpG sites

A linear regression model was used to identify sperm-specific DNA methylation signals on the 6,288 samples (371 of which were sperm). The linear models were performed on 452,626 CpGs using the lmFit function from the limma R package and included age and array type as covariates. The data was not normalised because global large-scale differences between somatic tissues and sperm were expected, and because of the high number of different types of samples included.

With Bonferroni correction, a total of 156,654 genome-wide significant CpG sites were identified as differentially methylated between sperm and somatic tissues (Figure 5-25).

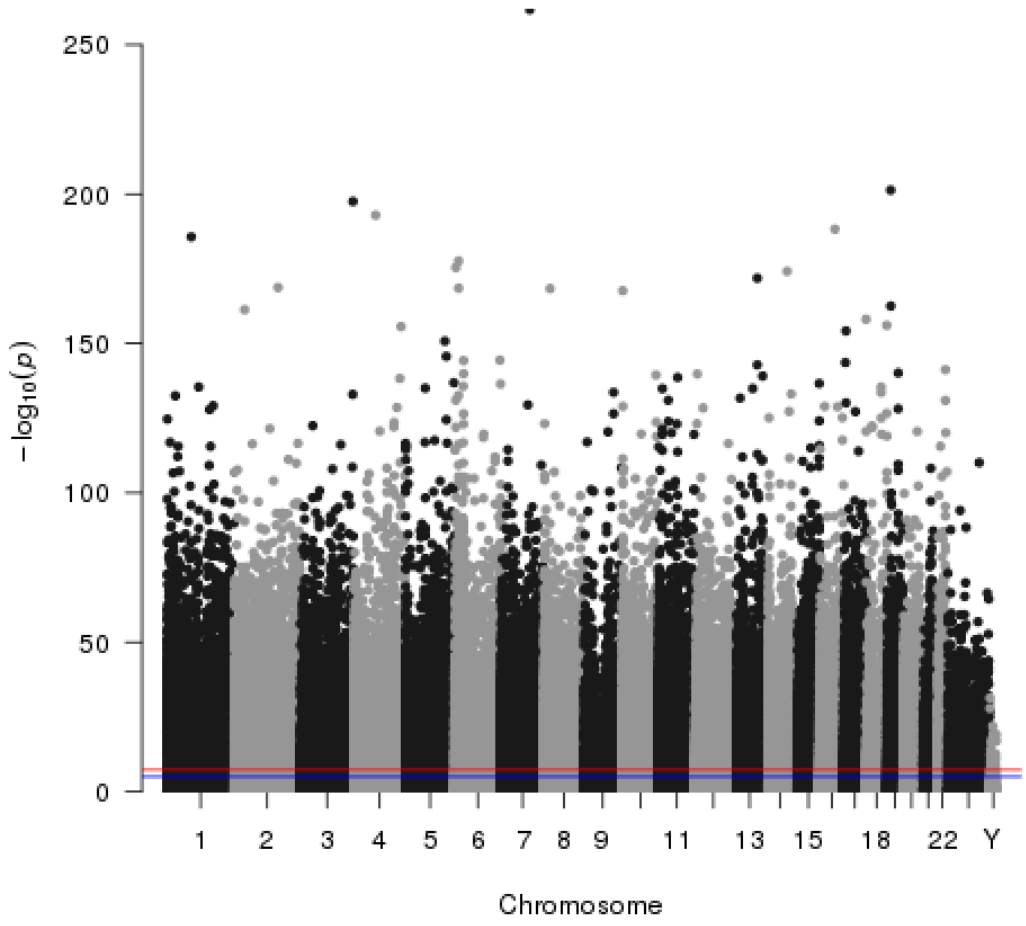


Figure 5-25. Manhattan plot of CpG sites that are differentially methylated between sperm and somatic tissues on GEO
The red line represents a Bonferroni adjusted significance level and the blue line represents an FDR adjusted significance level.

5.9.2 Gene Ontology pathway analysis

The 156,654 genome-wide significant CpG sites were annotated using the Gene Ontology (GO) database and enrichments for specific biological pathways were calculated (205). Enriched gene ontology terms from the Bonferroni significant CpGs were identified separately for both hypermethylated and hypomethylated CpGs using the gometh function from the missMethyl R package (207). No GO terms were found to be significantly enriched for hypomethylated CpG sites. However, six GO terms were identified as significantly enriched amongst hypermethylated CpG sites. These six terms are detailed in Table 5-4.

GO ID	GO TERM	N	DE	P.DE
GO:0000981	DNA-binding transcription factor activity, RNA polymerase II-specific	1537	835.2	3.72E-14
GO:0045944	Positive regulation of transcription by RNA polymerase II	975	559	1.88E-11
GO:0001228	DNA-binding transcription activator activity, RNA polymerase II-specific	390	247.3	1.38E-09
GO:0000978	RNA polymerase II proximal promoter sequence-specific DNA binding	379	233	1.82E-07
GO:0003700	DNA-binding transcription factor activity	480	278.8	4.30E-06
GO:0008134	Transcription factor binding	264	164.5	1.38E-05

Table 5-4. Gene ontology terms that were significantly enriched for hypermethylated CpG sites when comparing sperm to somatic tissues
GO = Gene Ontology, N = number of genes in the GO term, DE = number of genes that were differentially methylated, P.DE = p-value for over-representation of the GO term

As seen in Table 5-4, the GO terms that were significantly enriched for hypermethylated CpG sites all related to transcriptional regulation, either through TF binding or RNA polymerase activity. In very general terms, DNA hypermethylation, especially of gene promoters, is associated with transcriptional repression (266). Spermatozoa represent highly specialised cells with the main function of delivering genetic material to the oocyte. They have previously been shown to contain a low number of RNA transcripts compared to somatic cells (265). The significant hypermethylation enriched GO terms supports this observation; that spermatozoal gene expression is characterised by overall transcriptional repression.

5.10 Discussion

To my knowledge, this study constitutes the largest genome-wide investigation of CpG methylation in matched sperm and blood samples performed to date. It also constitutes one of the largest genome-wide investigations of sperm from healthy, fertile males.

The main conclusion is that the DNA methylation profiles of sperm and blood are clearly distinct, with highly significant mean differences of methylation levels at multiple sites across the genome. Indeed, factors such as age and interindividual variation make a much smaller contribution to overall variation in DNA methylation than do differences between the two blood and sperm.

A small subset of CpG sites that displayed a high DNA methylation co-variation between the two tissues. However, closer analysis showed that methylation at these sites was most likely influenced by underlying genetic variation. Most of the identified genetic variants were directly in the CpG site itself, confounding the DNA methylation measurements. Beyond that, even genetic variants in the probe sequence are known to affect binding properties and therefore induce technical artefacts in DNA methylation quantification. The small number of sites (10 out of ~700,000) where no obvious genetic driver of methylation variability was identified are likely too few to be of value in studies where blood is needed as a surrogate tissue for sperm. These findings demonstrate that the use of blood DNA methylation as a proxy for sperm DNA methylation is inappropriate.

The results of this study are in line with similar studies of epigenetic co-variation. For example, whole blood has been found to be an inappropriate proxy tissue for understanding DNA methylation in several brain regions (368). It should be said, however, that the results of this study do not preclude the use of blood for, for example, identifying CpG biomarkers for conditions such as subfertility or other reproductive conditions. For example, if a robust and reproducible blood DNA methylation profile of subfertility is identified in blood, this could be a helpful test in

fertility evaluations without it necessarily reflecting the epigenetic profile of spermatozoa.

This study identified one CpG site, cg19357369, as differentially methylated in sperm and blood from obese versus lean males. This finding should be interpreted with caution as it requires replication. Interestingly, however, this CpG site has, to my knowledge, not been identified in EWASs of obesity/BMI when only blood samples have been analysed. Therefore, it could point towards the possibility of an obesity associated signature of spermatozoa. Of note, cg19357369 has previously been identified as differentially methylated in hepatic tissue from obese compared to lean males (377). Another study found that methylation at cg19357369 is significantly altered during human fetal brain development (378). This study analysed genome-wide CpG methylation in fetal brain samples from day 23 to 184 post-conception, and found that methylation at cg19357369 peaked in fetal brains around day 70 post-conception (378). cg19357369 is found in the north shore of the gene for a large intergenic non-coding RNA (lincRNA) called *RP11-503N18* (377). To my knowledge, this lincRNA has yet to be characterised in terms of biological function.

The wider coverage of the MethylationEPIC array compared to its predecessor, the HM450, allowed the present study to refute some earlier general characterisations of the sperm epigenome. For example, previous analyses have suggested that the sperm epigenome is overall hypomethylated compared to somatic tissues (168, 343). My results show that the sperm epigenome is overall slightly hypermethylated compared to peripheral blood. Previous findings regarding overall hypomethylation in sperm may have arisen from the bias towards assessing promoter methylation in the HM450. In line with this, the present study showed that CpG islands in spermatozoa are significantly hypomethylated compared to CpG islands in blood.

Among previous analyses of the human sperm methylome, not only studies using the HM450 have suggested that the human sperm methylome is relatively hypomethylated compared to somatic tissues such as components of blood (330). For example, Rakyan et al employed a combination of methylated DNA

immunoprecipitation (MeDIP) and a custom high-density microarray to profile 13 normal somatic tissues as well and sperm and placenta, finding that sperm was characterised by overall hypomethylation relative to somatic tissues (including B-cells, CD4+ T cells, CD8+ T cells, colon, liver and uterine tissue) (330). Again, the results study are likely to reflect the methodology used; the custom oligonucleotide array was heavily biased towards interrogating CpG sites in CpG islands, and, consistent with our findings, such regions are characterised by relative hypomethylation in spermatozoa (330).

Further, in the context of genomic regions characterised by hypomethylation in spermatozoa relative to somatic tissues, we identified that CpG island shores were hypomethylated in spermatozoa compared to the equivalent regions in components of blood. This finding is in line with a previous study that used whole genome shotgun bisulfite sequencing and identified that spermatozoa exhibits relative hypomethylation in regions adjacent to CpG islands compared to, for example, human embryonic stem cells (379).

It should be said, however, that despite the relatively wider coverage of the MethylationEPIC array compared to its predecessor, the MethylationEPIC array is still biased towards certain parts of the genome (most notably enhancer regions, RefSeq genes and CpG islands) and does not give an complete picture of genome-wide CpG methylation (199). It will only be with analysis whole genome methylation that conclusive statements can be made regarding whether one tissue exhibits overall hyper- or hypomethylation in relation to another.

The comparative analysis between DNA methylation in sperm and DNA methylation of >6,000 somatic tissue samples on the Gene Expression Omnibus (GEO) database supported previous studies showing that spermatozoal gene expression is characterised by overall transcriptional repression (265). This is to be expected in a tissue with highly specialised functional requirements and tightly packaged chromatin.

In the publication of the widely used DNA methylation age tool developed by Steve Horvath, the author specifically mentions that the tool does not provide an accurate estimated chronological age for sperm, which was consistent with our findings (196). This can most likely be explained by the fact that the tool was constructed using data from only 45 samples of semen in a total of 7,844 samples (0.6%) of different tissue types, whereas it used data from 4,180 blood-derived samples (53%) (Additional file 1 from (196)). The model specifically developed by Jenkins et al to predict chronological age from genome-wide sperm DNA methylation data more accurately reflected the chronological age of participants where age data was available (372).

DNA methylation age tools, such as the one developed by Horvath, have been shown to not only be able to accurately predict chronological age from CpG methylation of DNA samples from various tissues, but also (to an extent) predict the biological age of the tissue (196, 380, 381). In this context, the 'biological age' refers to changes associated with the functional decline of a tissue that naturally occurs with aging, but which can occur more rapidly in, for example, disease states. The biological age is thus different from the chronological age, which is wholly determined by the time elapsed since an individual's birth. In future studies, it would be interesting to see whether there are reliable associations between phenotypic traits, such as disease states, and accelerated rate of biological aging in spermatozoa.

5.10.1 Strengths and Limitations

Strengths of this study include the matching of sperm and blood samples, the use of the most comprehensive DNA methylation array available to date, and the inclusion of healthy, fertile males rather than individuals recruited in a reproductive medicine setting. The latter allowed me to comprehensively characterise the human sperm methylome in a manner that is likely to be generalisable to a large proportion of the male population. In terms of the GEO analysis, strengths included using the bigmelon analysis pipeline that allowed the inclusion of a large number (>6,000) of somatic tissue samples and thus greatly added power. Also, whilst the main part of this study focussed on comparing the DNA methylomes of sperm and blood specifically, those results could reflect leukocyte specific gene expression requirements, i.e. be less

valuable when trying to elucidate sperm-specific gene expression requirements. Thus, the addition of the GEO analysis improved the ability to identify sperm specific biological processes.

This study also has a number of limitations. Whilst having a large sample size compared to previous analyses of gamete epigenomes, the sample size is limited compared to EWASs of, for example, peripheral blood (106). In particular, the obesity cohort (n = 24) was likely too small to detect modest differences in sperm-blood methylation covariation between lean and obese males. The obesity cohort also included some participants that were overweight (BMI 25-30 kg/m^2) rather than obese (BMI >30 kg/m^2). This potentially contributed to the lack of significant findings. We were able to speculate as to the effects of genetic variants in CpG sites influencing our results, given trimodal methylation patterns and the presence of known SNPs in the CpG site. However, we did not have the actual genetic sequence of our subjects to verify this directly. It is likely however, that for the majority of CpG sites exhibiting significant correlation between blood and sperm methylation, genetic variation is the driver. In this regard, it is surprising that several thousand variants seem to have escaped previously compiled lists of genetic variants in probe sequences of 450K or EPIC arrays (87, 370, 371). While genetic variation in DNA methylation probes represents a known challenge in the interpretation of methylation array data, it nevertheless added a layer of difficulty in the identification of sites with a high methylation co-variation. In the case of the GEO analysis, limitations include the lack of normalisation of the datasets included in the analysis and that cross-hybridising and probes with high detection p-values were not filtered out.

5.10.2 Future Directions

In line with previous research, the present study demonstrated that the human sperm methylome is highly polarised towards high and low levels of methylation compared to blood (168). In future research, it would be interesting to functionally explore genomic regions that display high versus low levels of methylation in sperm compared to somatic tissues, e.g. by expression profiling.

This study also identified CpG sites where DNA methylation levels were significantly correlated, but at the same time displayed highly discordant DNA methylation levels between the two tissues. These included sites where, for example, the DNA methylation level in sperm was low and similar across individuals, but displayed a trimodal distribution pattern in blood. It is possible that tissue specific transcription factors underlie such differences. In this regard, it is interesting to consider that all of the GO terms that were significantly enriched for hypermethylated CpG sites in sperm compared to somatic tissues related to transcriptional regulation, mostly via TF binding. Better characterisation of sperm-specific transcriptional regulation and CpG sites that show highly discordant methylation levels between sperm and somatic tissues would yield insights into which biological processes are important for germ cell development and function. This, in turn, could aid the understanding of fertility, embryogenesis and the potential for germ cells to respond to environmental and physiological change.

Future investigations should avoid using blood as a proxy tissue for analyses of sperm DNA methylation. They should also better characterise the methylation 'escape variants', i.e. CpG sites that appear to escape the global DNA demethylation that occur shortly following fertilisation and during gamete development (128). Such research would be more likely to yield informative insights into the potential for acquired traits to influence the next generation than would additional studies of surrogate tissues such as blood.

In the context of escape variants, these appear to be particularly enriched for in repeat elements of the genomes, including retrotransposons (128). Further, there appears to be a correlation between how evolutionarily young a particular retrotransposon is, and its propensity for resisting demethylation (128). For example, a considerable fraction of loci within SINE-variable number of tandem repeats-Alu elements (SVAs), which are relatively evolutionarily young and active, appear to resist the demethylation process that other retrotransposons undergo in during, typically, weeks 5.5-9 of human primordial germ cell development (128). The authors suggest that, together with other methods of regulation of gene expression, e.g. by

transcription factor binding, this resistance to demethylation in evolutionarily young retrotransposons may limit potentially hazardous retrotranspositions of SVAs in the germline (128).

This study identified a number of specific CpG sites that would be interesting to take forward for further investigation. For example, 42 CpG sites were identified where there was a significant *negative* correlation between DNA methylation in blood and DNA methylation in sperm (section 5.7.2). Such sites have previously been characterised when comparing genome wide CpG methylation in blood with that of different brain regions (368). To our knowledge, however, this is the first study to identify such sites when examining blood and sperm, and it is fascinating to hypothesise about the potential underlying mechanism. For example, could there be an environmental or physiological trait that influences both sperm and blood DNA methylation but in opposite directions?

5.11 Summary

In this study I investigated genome-wide CpG methylation in 92 matched sperm and blood samples using the Illumina MethylationEPIC array. DNA methylation profiles of the two tissues were found to be highly discordant, with few CpG sites confidently demonstrating a methylation correlation between the tissues. Results from this study confirm the importance of using disease relevant tissues in epigenomic investigations and question the validity of previous analyses where blood has been used as a proxy tissue for sperm DNA methylation.

This study also identified a number of particular CpG sites that would be interesting to take forward in future investigations. These include CpG sites that display negatively correlated DNA methylation between sperm and blood, and CpG sites that, despite being significantly correlated, show near complete methylation in one of the tissues but are near complete lack of methylation in the other. The study identified one CpG site, cg19357369, as differentially methylated sperm and blood from obese versus lean males. This CpG site has not previously been identified in

EWASs of obesity performed on blood only, and would be interesting to further characterise.

A comparison of DNA methylation in sperm to that of a wide range of somatic tissues suggested that there is an overall tendency towards transcriptional repression in spermatozoa.

Characterisation of the human sperm epigenome has important implications for understanding fertility, embryogenesis and the potential for germ cells to respond to environmental and physiological change. The findings presented in this study contribute towards that aim.

Chapter 6
Obesity Associated DNA Methylation Profiling in Blood

6.1 Introduction

Obesity and its related metabolic syndrome constitute major public health problems globally (13). In 2016, 39% of all adults globally, more than 1.9 billion people, were overweight (BMI >25 kg/m^2), of whom one-third were obese (BMI >30 kg/m^2) (2). Once obesity is acquired, the probability of achieving and maintaining a normal body weight is low (17). Despite decades of rising awareness of the morbidity, mortality and financial implications of the obesity epidemic, no country has successfully sustained a reversal of current trends (16).

Obesity and overweight most commonly arise as a result of excessive calorie intake and a sedentary lifestyle (2). However, obesity is a multifactorial condition that involves a complex interplay of genetic, gene-environment, environmental and behavioural factors. Although some genetic variants are robustly associated with the development of obesity, these can only explain a minority of cases (29). Further, the rising prevalence of obesity has been too rapid to be explained by genetic factors alone (382). Therefore, a major research focus has been to conduct increasingly large epigenome wide association studies (EWASs) of obesity (discussed in detail in section 1.8.1). However, despite relatively large sample sizes, few of the CpG sites identified in these EWASs have been replicated across studies. There is therefore value in validating the previously identified CpG sites to assess whether they are replicable across cohorts.

The identification of a robust, reproducible DNA methylation profile associated with obesity would improve our understanding of the pathogenesis of the metabolic syndrome, identify biomarkers for disease progression, and therapeutic targets.

In the present study I validated a set of 192 previously identified obesity associated CpG sites in a novel cohort of 96 obese men (BMI > 30 kg/m^2) compared with 96 lean men (BMI 18-25 kg/m^2). This study forms the first stage of a larger project that aims to generate a robust, reproducible obesity-associated DNA methylation profile using peripheral blood from a total of 1000 lean and 1000 obese males (MRC reference code MR/P011799/1; title *'Paternal obesity-associated DNA methylation: an*

investigation into its reproducibility, reversibility and association with fetal growth restriction').

6.1.1 Declarations

Peripheral blood samples were collected as part of the Dad's Health Study by myself and Anna Greco, research midwife, at UCLH. Peripheral blood samples from the Iowa cohort were collected by Dr Donna Santillan at the Women's Health Tissue Repository, University of Iowa Health Care. DNA extraction from the Dad's Health and Iowa cohort blood samples was performed by myself and Anna Greco. CpG sites from previous obesity EWASs were identified by Dr Sarah Marzi at Queen Mary University London (QMUL), who also designed the primers for the multiplex bisulfite PCR sequencing microfluidics-based assay. Primers were tested for efficacy by Adrian Signell and Dr Michelle Holland at Kings College London. Bisulfite conversion of blood samples was performed by Theodoros Xenakis at QMUL. The multiplex bisulfite-PCR-sequencing microfluidics-based assay, library preparation and subsequent Next Generation Sequencing (NGS) was performed at the Genome Centre Facility at QMUL. DNA methylation data preprocessing and analysis was performed by Dr Sarah Marzi at the Blizard Institute, QMUL.

6.2 Hypothesis

Differential DNA methylation of CpG sites associated with obesity from previous EWASs can be replicated in a new cohort of obese men.

6.3 Specific Objectives

1. To identify which CpG sites previously identified as associated with obesity in EWASs are significantly associated with obesity in a novel cohort of 96 lean and 96 obese males.

6.4 Methods

6.4.1 Study design and population

Ethical approval was granted from the South East Coast - Surrey Research Ethics Committee on 28 September 2015 (REC reference number 15/LO/1437, IRAS project ID 164459). The study was also registered with the UCLH Joint Research Office (Project ID 15/0548). All participants provided written, informed consent. Ethical permission for the transfer of peripheral blood samples collected in Iowa was provided via a materials transfer agreement (MTA-17-252).

CpG sites were identified from six EWASs of BMI and used to generate specific primers as described in Chapter 2 section 2.14-2.15 (101-106). CpG sites were also included from a study investigating the association between obesity associated SNPs and methylation of nearby CpG sites (212). Smoking and inflammation are important covariates in studies of obesity, as they are known to influence DNA methylation and can confound results. Therefore, CpG sites where methylation levels have been shown to be influenced by smoking and inflammation (as measured by C-reactive protein, CRP) were included as controls for these covariates. The smoking associated sites were identified from a meta-analysis of EWASs of smoking (195). The CRP associated sites were identified from a meta-analysis of EWASs of serum C-reactive protein (213). Blood represents a heterogenous tissue containing numerous different types of leukocytes, which can lead to confounding in EWASs using blood (382). In the present study, this was accounted for by using highly cell type specific methylation signatures as proxies for cell type composition. The blood cell composition control sites were identified from (214).

Primers were evaluated for specificity and efficacy as described in Chapter 2 section 2.14.1. 192 of the primers that performed well in the evaluation assay were selected to be included in this study. The 192 selected primers consisted of 140 that targeted CpG sites identified from obesity EWASs, 14 that targeted CpG sites associated with smoking or CRP, 27 that targeted CpG sites associated with blood cell composition, and 11 that targeted CpG sites associated with SNPs of obesity.

Peripheral blood samples were collected as part of the Dad's Health Study at University College London Hospital (UCLH) between May 2016 and March 2019 as described in detail in Chapter 2 sections 2.3-2.6. DNA was extracted as described in Chapter 2 section 2.8 and bisulfite converted as described in Chapter 2 section 2.11.

Quantification of CpG methylation in the 192 regions of interest in bisulfite converted genomic DNA from 96 lean and 96 obese males was performed using a multiplex bisulfite-PCR-sequencing microfluidics-based assay, hereafter referred to as the 'Bis-PCR-Seq assay'. This is described in detail in Chapter 2 section 2.15. Briefly, regions of interest in bisulfite converted genomic DNA were amplified using the 48.48 layout on the Fluidigm® C1 system (Fluidigm®, USA) (218, 383). Next Generation Sequencing (NGS) libraries were generated from the amplicons using the same kit, including 4 µl of Access Array Barcode Library Primer and 1 µl of PCR product diluted 1:100. Libraries were sequenced on an Illumina MiSeq sequencer (150 bp, paired-end). The output data was provided as FastQ files, which are text files containing the target sequence together with a quality score (384).

6.4.2 Statistical analyses

Phenotype analyses were carried out using RStudio version 1.1.456. Sequencing reads were aligned to a bisulfite converted reference genome, the GRCh38.p13, using Bismark (385, 386). The 5mC level at each CpG site was calculated using a customised python script, calculating the methylation value β as the ratio of methylated reads over the total number of reads at each covered CpG site (387).

Measurements of BMI, waist circumference, systolic and diastolic blood pressure, total cholesterol as well as HDL and LDL cholesterol were found to be approximately normally distributed (Shapiro-Wilk test $p>0.10$), and are therefore summarised as mean values and standard deviations of the mean. The means of these measurements were compared between the lean and obese cohorts using a Welch t-test. The distribution of measurements of fasting glucose, C-reactive protein and triglyceride levels as well as calculated HOMA-IR values was found to be skewed (Shapiro-Wilk test $p<0.10$), and are therefore summarised as median values and

interquartile ranges. In these cases, a Wilcoxon rank sum test was used to compare median values between the cohorts.

6.5 Results

6.5.1 Phenotype profiles of study participants

Phenotype characteristics of lean (BMI 18-25 kg/m^2) and obese (BMI >30 kg/m^2) participants are presented in Table 6-1.

There were clear, significant differences in metabolic characteristics between lean and obese participants (Table 6-1). Measurements of BMI, waist circumference, systolic and diastolic blood pressure, total- HDL- and LDL-cholesterol, fasting glucose and insulin, HOMA-IR, HOMA2-IR, CRP and triglycerides were all significantly different between the two groups ($p < 0.001$ for all but fasting glucose where $p = 0.001$ i.e. all statistically significant).

	Lean	Obese	Reference Range	p
n	96	96		
Age, years. Mean (SD)	35.8	37.0	N/A	0.232
BMI, kg/m^2. Mean (SD)	23.2 (1.2)	33 (3.3)	18.5–24.9	<0.001
Waist circumference, cm. Mean (SD)	82.3 (10.5)	110 (9.6)	< 94 cm	<0.001
SPB, mmHg. Mean (SD)	119 (13)	131 (16)	90 - 120	<0.001
DPB, mmHg. Mean (SD)	76 (9)	84 (12)	60 - 80	<0.001
Total cholesterol, mmol/L. Mean (SD)	4.8 (1)	5.3 (0.9)	< 5.0	<0.001
HDL cholesterol, mmol/L. Mean (SD)	1.6 (0.6)	1.2 (0.3)	0.9-1.5 (males)	<0.001
LDL cholesterol, mmol/L. Mean (SD)	2.8 (0.9)	3.2 (0.8)	< 3.5	<0.001
Fasting glucose mmol/L. Median (IQR)	4.7 (0.5)	4.9 (0.5)	3.9-5.8	0.001
Fasting insulin, mIU/L. Median (IQR)	5.8 (2.9)	13.0 (10.5)	2.6-24.9	<0.001
HOMA-IR. Median (IQR)	1.2 (0.6)	2.7 (2.7)	Usually ≤ 2.0	<0.001
HOMA2-IR. Median (IQR)	0.8 (0.4)	1.7 (1.4)	Usually ≤ 1.8	<0.001
CRP, mg/L. Median (IQR)	0.6 (0.4)	1.6 (2.3)	0-5.0	<0.001
Triglycerides, mmol/L. Median (IQR)	0.9 (0.5)	1.6 (1.3)	<2.3	<0.001

Table 6-1. Phenotype characteristics of study participants.
Reference ranges are derived from the UCLH Clinical Biochemistry Test Information sheet available from (230). The reference range for HOMA-IR is derived from (231). The reference range for HOMA2-IR is derived from (232). The reference range for waist circumference is derived from (179). The reference ranges for blood pressure are derived from (233). SD = Standard Deviation, IQR = interquartile range, BMI = Body Mass Index, SBP = Systolic Blood Pressure, DBP = Diastolic Blood Pressure, HOMA-IR = Homeostatic Model Assessment of Insulin Resistance, CRP = C-Reactive Protein, HDL = High Density Lipoprotein, LDL = Low Density Lipoprotein.

6.5.2 Quality Control of Sequencing Data

Illumina MiSeq sequencing data was provided as FastQ files. FastQC quality control profiles were generated for all FastQ files and no large abnormalities or low quality data was detected for any files (an example of FastQC files of the Illumina MiSeq sequencing data is provided in Appendix 5) (384, 388).

6.5.3 Aligning of amplicons to a reference genome

Reads were aligned to an *in silico* bisulfite converted reference genome (GRCh38.p13) using Bismark (385, 386). In the bisulfite converted reference genome, non-CpG cytosines are replaced with thymines (T), whereas CpG cytosines remain as cytosines (C) (383). A phred score of 10, equivalent to a base call accuracy of 90%, was set as a minimum quality score for amplicons to align to the reference genome (389). CpG methylation was quantified by observing the base calls at the CpG sites in the mapped reads; unmethylated CpG sites in the target sequences would have T whereas methylated CpG sites in the target sequences would have C in the mapped reads. Methylation levels were expressed as percentages, ranging from 0 (completely unmethylated) to 100 (fully methylated). For each CpG site, the methylation level in the original sample, expressed as a percentage, was determined as follows:

$$\text{Methylation level } (\beta) = \frac{\text{C}}{C + T}\ x\ 100$$

A relatively low mapping efficiency was observed when aligning the reads to the reference genome (mean 19.4%, median 20%, range 8.4-24.9%). In other words, most of the reads mapped non-specifically and could not be analysed adequately. No participants were excluded due to low mapping efficiency. An average read coverage of 20 was used as a minimum cut-off for each CpG site. This meant filtering of 113 amplicons, leaving 79 amplicons containing CpG sites of interest for further analyses. Of these 79, 64 were from obesity EWASs, 3 were control probes for smoking or CRP, 6 were control probes for blood cell type composition and 6 were for CpG sites associated with obesity associated SNPs (Figure 6-1).

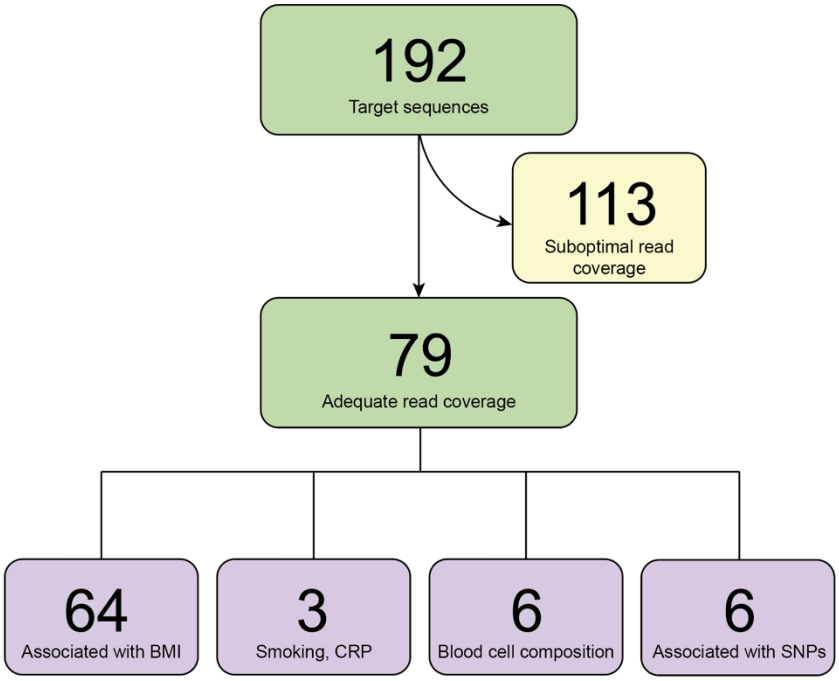


Figure 6-1. Amplicons excluded based on low read coverage

Several of the amplicons contained multiple CpG sites in addition to the specific one of interest. This meant that a total of 441 CpG sites were included in downstream analyses.

6.5.4 Principal component analysis

Principal component analysis (PCA) of the 192 samples revealed no overall difference between samples according to the cohort they had been collected from, i.e. UCLH or Iowa (Figure 6-2). It did identify three samples (two from the Iowa cohort and one from the UCLH cohort) as outliers. These were removed from further downstream analyses. Thus, 189 samples remained.

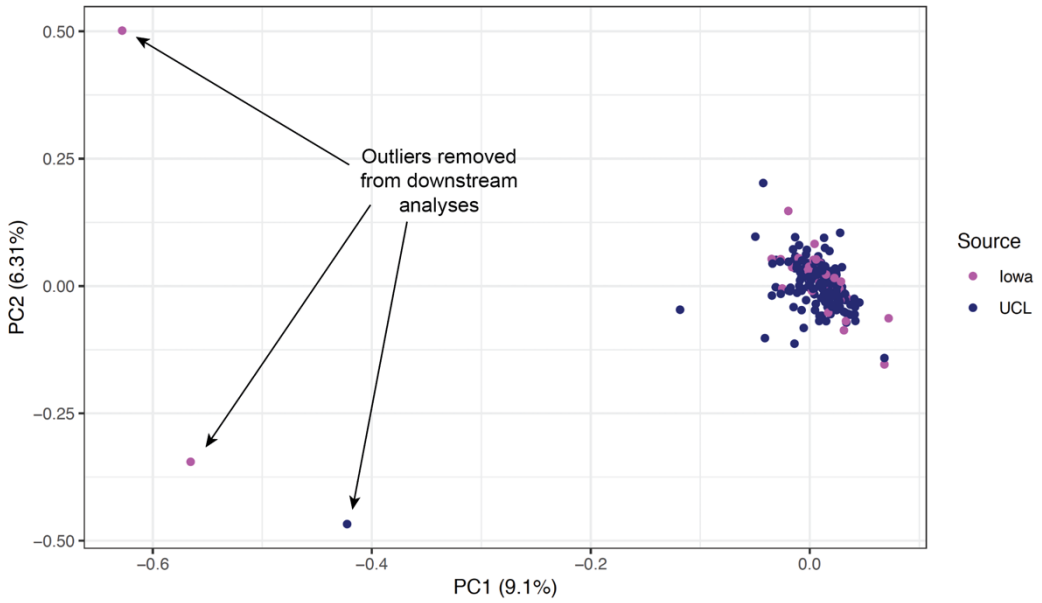


Figure 6-2. Principal component analysis of the 192 samples included in the study of validating obesity associated CpG sites
The three samples indicated by the arrows were identified as outliers and were removed from downstream analyses. PC = Principal Component

6.6 Association between CpG methylation and Obesity Status

Different approaches were used to test for an association between CpG methylation at the 70 sites of interest (64 from obesity EWASs and 6 associated with obesity SNPs) and BMI.

The main analysis model was a linear regression with methylation as the outcome regressed onto BMI, age, blood cell composition (B cells, CD4+ T cells and monocytes), smoking and CRP. In this model, one CpG site was picked randomly as a proxy for each of the covariates for which multiple associated CpG sites were

available (i.e. smoking, monocytes and CD4+ T cells). These were the only covariates which had CpG sites with sufficient read coverage. With FDR correction for multiple testing, no CpG sites were identified as significantly associated with BMI in this model. This model most closely reflects the methods used in the EWASs from which CpG sites were selected, and is therefore the focus of the results from this study (hereafter referred to as the 'main' linear model). Results for the 70 informative probes (64 from obesity EWASs and 6 that are associated with obesity CpGs) are presented in Table 6-2, where they are ranked according to the p value for their association with BMI.

Probe	p (linear model)	p (main linear model, FDR corrected)
cg07037944	0.001225299	0.06758115
cg26651978	0.00193089	0.06758115
cg07960624	0.028959025	0.53335775
cg00431050	0.030477586	0.53335775
cg00108715	0.043737376	0.61232326
cg03078551	0.092536946	0.67541442
cg14264316	0.152834633	0.67541442
cg15442888	0.15059464	0.67541442
cg03957124	0.154380438	0.67541442
cg15357118	0.096551232	0.67541442
cg09363892	0.075239381	0.67541442
cg20507228	0.127862974	0.67541442
cg25570328	0.117257923	0.67541442
cg08877257	0.15074389	0.67541442
cg13840239	0.14040811	0.67541442
cg03433986	0.152859854	0.67541442
cg12917475	0.210549984	0.71687174
cg07504977	0.235146089	0.71687174
cg00916899	0.217534508	0.71687174
cg25217710	0.235206539	0.71687174
cg25096107	0.236494038	0.71687174
cg16395997	0.234860561	0.71687174
cg19750657	0.245784598	0.71687174
cg01101459	0.243367995	0.71687174
cg08857797	0.385964057	0.76683972
cg10508317	0.354063565	0.76683972
cg05628049	0.274293957	0.76683972
cg10717869	0.336288407	0.76683972

cg01677628	0.312000927	0.76683972
cg01798813	0.288047918	0.76683972
cg26257082	0.305562691	0.76683972
cg00994936	0.370362156	0.76683972
cg00834536	0.368435589	0.76683972
cg10734665	0.378845657	0.76683972
cg15497724	0.363650291	0.76683972
cg22143698	0.394374716	0.76683972
cg09349128	0.482897241	0.84507017
cg13997435	0.473347066	0.84507017
cg04924511	0.455350281	0.84507017
cg02008402	0.478693773	0.84507017
cg26357885	0.500347584	0.85425197
cg07728579	0.531564474	0.88594079
cg00634542	0.627067371	0.89581053
cg03327570	0.618423436	0.89581053
cg01881899	0.620896593	0.89581053
cg11152384	0.607840439	0.89581053
cg27269962	0.596640489	0.89581053
cg23172671	0.619094684	0.89581053
cg00489954	0.591012002	0.89581053
cg09109383	0.685979872	0.95301219
cg09222732	0.702631133	0.95301219
cg24145109	0.707951914	0.95301219
cg11376147	0.787153269	0.96115174
cg13084458	0.775574576	0.96115174
cg16611584	0.815517269	0.96115174
cg16721489	0.771621426	0.96115174
cg24824917	0.764482708	0.96115174
cg05149343	0.837575085	0.96115174
cg13010621	0.735506089	0.96115174
cg22950899	0.826453951	0.96115174
cg04816311	0.799271833	0.96115174
cg04557677	0.853213285	0.96330532
cg17560136	0.911985835	0.98213859
cg11660018	0.892933475	0.98213859
cg23576855	0.903057926	0.98213859
cg18217136	0.930732814	0.98714086
cg06876354	0.973260779	0.99181011
cg23417875	0.971493686	0.99181011
cg09956615	0.977641392	0.99181011
cg14020176	0.992235325	0.99223532

Table 6-2. *(Previous page)* Results from the main linear regression model of DNA methylation regressed onto BMI, age, blood cell composition, smoking and CRP
Of the 70 probes in the table, 64 were identified from obesity EWASs and 6 from a study of CpG sites associated with obesity SNPs. The CpG sites are ranked from lowest to highest FDR-adjusted p value.

As shown, no CpG site reached assay-wide significance after multiple testing correction. There were, however, two sites which were suggestive of an association with BMI. These were cg07037944 (p = 0.068) and cg2665197 (p = 0.068) (Figure 6-3). It is possible that a larger study cohort would have identified these as significantly associated with BMI. All other sites had FDR corrected p values for an association with BMI >0.5.

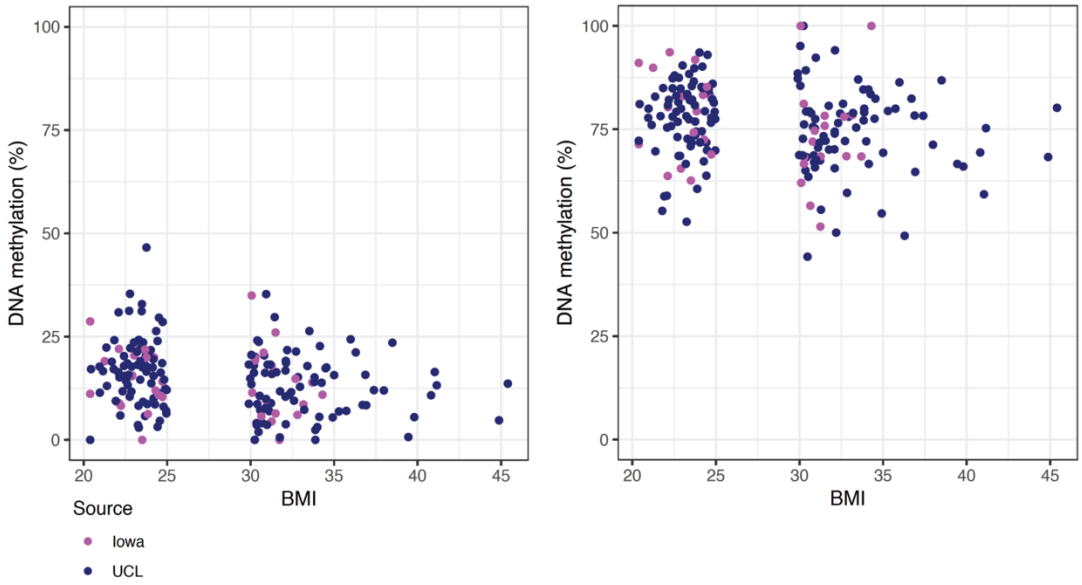


Figure 6-3. Scatter plots for methylation levels in the two CpG sites with the lowest p value for an association with BMI in the main linear regression model
a. cg07037944 (p = 0.068)
b. cg26651978 (p = 0.068)

Another linear model that included all the CpG markers for the covariates, i.e. one for B cells, two for CD4+ T cells, three for monocytes, one for CRP and two for smoking was performed. It was first confirmed that the CpG sites selected were independently associated with the covariates, for example that the two CpG sites associated with smoking did not correlate with each other. This model did not identify any CpG sites as significantly associated with BMI after multiple testing correction.

Finally, a t-test for methylation level differences between the lean and obese participants was performed. With FDR correction for multiple testing, no CpG site

was identified as significantly associated with obesity. The p values for the two CpG sites identified as suggestive of an being associated with BMI in the main linear regression model were p = 0.201 for cg07037944 and p = 0.352 for cg26651978. For all the other 68 CpG sites of interest, FDR corrected p values for an association with BMI exceeded 0.5 (Appendix 4).

6.7 Comparison of results from the Bis-PCR-Seq assay with results from the original EWASs of BMI

Given the non-replication of previously reported results, a dependent sample sign test was used to investigate whether the direction of methylation change (i.e. higher or lower in relation to BMI) in the Bis-PCR-Seq assay was correlated to that observed in the original obesity EWASs. This analysis did not find that the direction of methylation change was significantly correlated (p = 0.10, data not shown).

Next, it was investigated whether the change in methylation per unit of BMI observed in obesity EWASs was correlated with the change in methylation per unit of BMI observed in the Bis-PCR-Seq assay. To this end, the raw methylation change per unit change of BMI was extracted from the obesity EWASs for the 70 informative probes (64 from obesity EWASs and 6 associated with obesity SNPs) or transformed when necessary. This data was available in all but one study, where methylation levels were reported as M values, i.e. the log2 ratios of the intensities of methylated probe versus unmethylated probe in the Illumina methylation arrays, rather than beta values (105, 390). The mean of these raw beta value changes was calculated for each of the 50 probes for which changes in beta value of methylation were available. This mean value was then assessed for correlation with the methylation change beta value observed in the Bis-PCR-Seq assay (Figure 6-4). Using Pearson's product-moment correlation, effect sizes were found to be significantly correlated (R^2 = 0.14, p = 0.006). In brief, this means that the change in methylation per unit of BMI observed in obesity EWASs was correlated with the change in methylation per unit of BMI observed in the Bis-PCR-Seq assay.

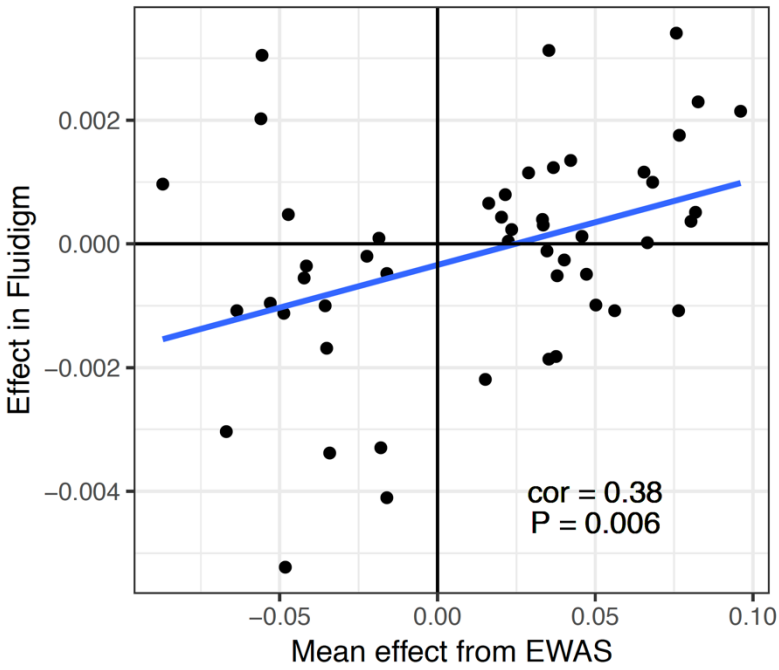


Figure 6-4. Comparison of mean effects in obesity EWASs to the effect observed in the Bis-PCR-Seq assay ('Fluidigm')
The black markers represent the 70 informative sites. The x axis shows the mean methylation effect observed in obesity EWASs and the y axis shows the effect observed in the Bis-PCR-Seq assay ("Fluidigm").

6.8 Comparison of results of the Bis-PCR-Seq assay to the MethylationEPIC array

Given the very limited replication of previous results observed in this study, we next investigated the characteristics and quality of the Bis-PCR-Seq data against a comparable sample of whole blood DNA methylation profiles from lean and overweight/obese men profiled on the Illumina EPIC array. Thus, it was investigated whether results obtained using the MethylationEPIC array (Chapter 5) correlated with those obtained using the Bis-PCR-Seq assay. DNA MethylationEPIC data from 21 lean (BMI 18-25 kg/m^2) and 22 overweight/obese (BMI> 25 kg/m^2) participants was included in this analysis. Measurements of BMI, waist circumference, fasting glucose, insulin resistance (HOMA-IR and HOMA2-IR) and CRP were significantly different between the cohorts of lean and overweight/obese participants ($p < 0.05$).

All the obesity EWASs from which CpG sites were selected were performed using the HM450. The MethylationEPIC array includes >90 % of the CpGs from the HM450 (87).

60 out of the 70 informative sites from the Bis-PCR-Seq assay were represented on the MethylationEPIC array.

The correlation of DNA methylation values obtained using these two methods is presented in Figure 6-5. As shown, there was a high and significant correlation between the median methylation levels obtained using the Bis-PCR-Seq assay and the MethylationEPIC array (R^2 = 0.90, p = 2.2 x 10^{-16}).

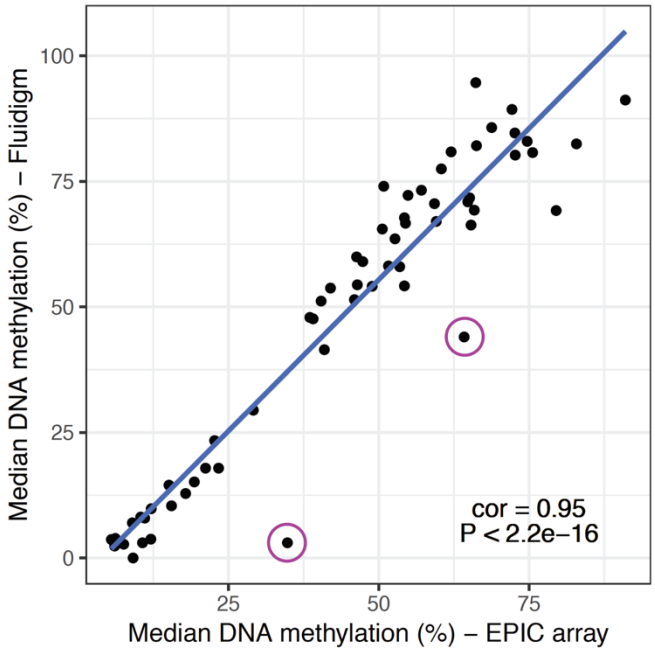


Figure 6-5. Comparison of results from the Bis-PCR-Seq assay ('Fluidigm') and the MethylationEPIC array
The black markers represent the CpG sites that were included in the Bis-PCR-Seq assay as well as being present on the MethylationEPIC array (n = 60). The two CpG sites with the highest difference in median methylation levels are circled. They correspond to cg04924511, which has a median methylation difference of 32% between the Bis-PCR-Seq assay and the MethylationEPIC array, and cg00489954, with a 29% median methylation difference.

It was clear that the results from the Bis-PCR-Seq assay overall reflected those obtained using the MethylationEPIC array. It was then assessed whether the methylation levels at the 60 probes individually showed the same direction of association with BMI when comparing results from the Bis-PCR-Seq assay to those obtained using the MethylationEPIC array. This yielded mixed results, with 30 of the 60 sites showing the same direction of association with BMI on the Bis-PCR-Seq assay and the MethylationEPIC array. Bis-PCR-Seq assay and MethylationEPIC data for the

two CpG sites which had the lowest p values in the main linear regression model of methylation regressed onto BMI, age, blood cell composition, smoking and CRP (Table 6-2) are visualised in Figure 6-6.

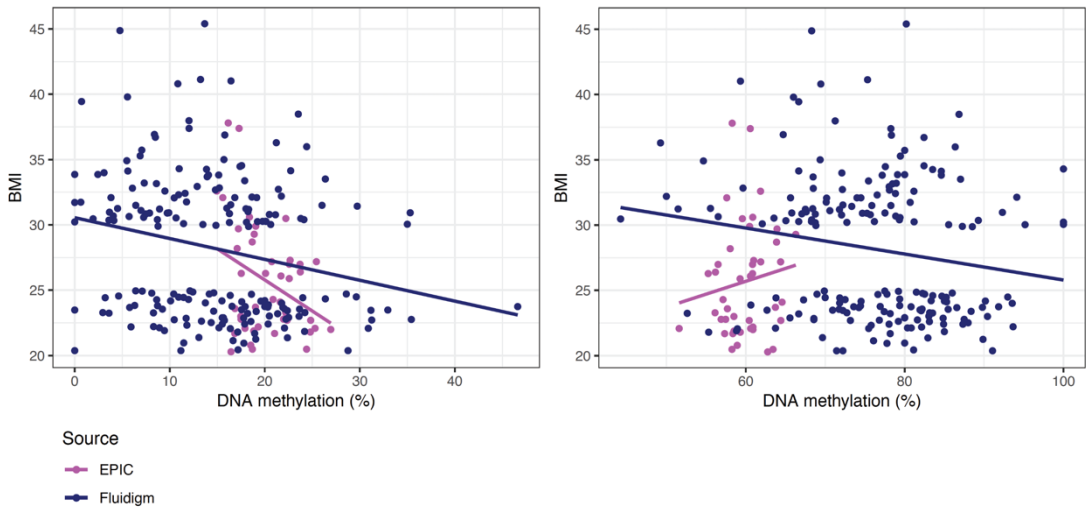


Figure 6-6. Comparison of DNA methylation at specific probes in the Bis-PCR-Seq assay and MethylationEPIC array
a) cg07037944
b) cg26651978

As exemplified in Figure 6-6, the variability of methylation results at individual CpG sites was considerably higher in the Bis-PCR-Seq assay compared to those measured using the MethylationEPIC array. For example, at probe cg07037944, the methylation levels vary between ~0-47% when measured using the Bis-PCR-Seq assay, but vary between ~15-26% when measured using the MethylationEPIC array. The higher methylation variability in the Bis-PCR-Seq assay was observed for all of the 60 CpG sites that were analysed in the Bis-PCR-Seq assay and represented on the MethylationEPIC array. This effect could, in part, be related to the level of read coverage of the individual CpG sites. Indeed, there was a significant correlation between the average coverage of each CpG site and the variability of methylation levels, measured as standard deviations (R^2 = 0.14, p = 8.4 x 10^{-4}) (Figure 6-7). Other potential explanations for the higher variability in the Bis-PCR-Seq assay include PCR-related artefacts.

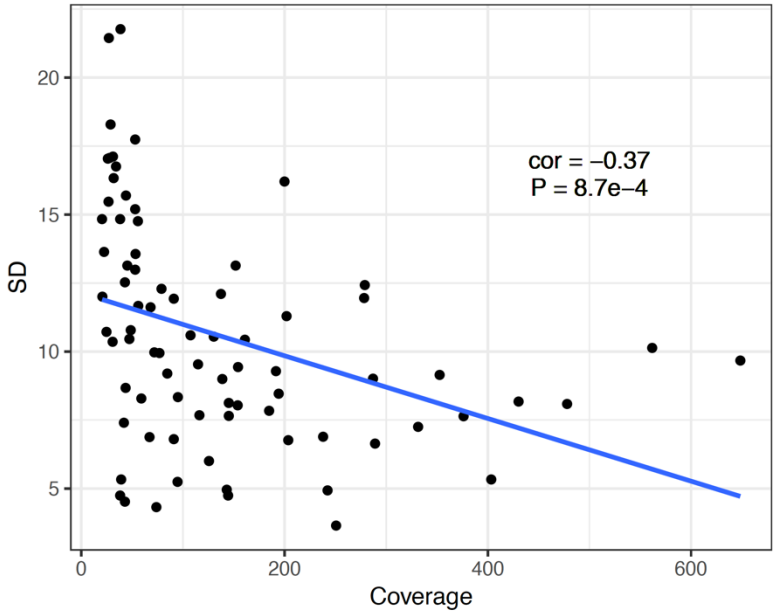


Figure 6-7. Correlation between average coverage of CpG sites and variability (expressed as standard deviation, SD) in the Bis-PCR-Seq assay

The black markers represent the 79 informative CpG sites that passed QC in the Bis-PCR-Seq assay. SD = Standard Deviation

6.9 Discussion

This study formed the first part of a large scale analysis aimed to identify a robust, replicable DNA methylation signature of obesity. Whereas the present study included ~200 participants, the larger analysis will include ~2,000 participants. Therefore, the results should not be seen as conclusive. The present study was nevertheless an important step in identifying suggestive CpG sites, evaluating the methodology and informing the next stages of the large scale project.

In the present study, despite comparing well-characterised obese and lean men, none of the previously identified CpG sites were replicated as significantly associated with obesity/BMI. Two CpG sites, cg07037944 and cg26651978 were suggestive of an association with BMI in a linear regression model that controlled for age, blood cell type composition, smoking and inflammation (FDR corrected p = 0.068 for both sites).

The CpG site cg07037944 is annotated to the gene Death Associated Protein Kinase 2 (*DAPK2*) (104). The CpG site is, however, not in the 5' promoter of the gene, where it would have been expected to have the highest likelihood of influencing expression of *DAPK2*, but is intragenic. The role of cg07037944 in the regulation of *DAPK2* should therefore be interpreted with caution. In either case, cg07037944 was identified as associated with BMI in two large scale EWASs of BMI, one of which was the reason for its inclusion in this analysis and the second published after the time of identifying CpG sites for the present project (104, 106). In addition, cg07037944 was associated with BMI change (when comparing BMI at age 18-21 and BMI in middle age) in a twin family study (391). *DAPK2* is a member of a family of kinases with cellular functions related to apoptosis, autophagy and immune system functioning (392, 393). Inhibition of *DAPK2* has been shown to reduce the inflammatory response to infection in a murine model (393). Obesity is a condition characterised by chronic, low-grade inflammation, which contributes to related conditions such as insulin resistance (394). Alterations in the function of *DAPK2* may therefore play a role in the pathogenesis of obesity.

The CpG site cg26651978 is annotated to the gene Lectin Galactoside-Binding Soluble 3-binding Protein (*LGALS3BP*) (104). An important caveat to mention here, however, is that cg26651978 is 3' of *LGALS3BP* and therefore, although *LGALS3BP* is the gene nearest to the CpG site, it may not actually be involved in its regulation. Several EWASs have, however, identified CpG sites annotated to *LGALS3BP*, as associated to BMI and related traits such as waist circumference (102, 106, 107). In one of these studies, CpG sites annotated to *LGALS3BP* were also found to be associated with BMI in adipose tissue (106). *LGALS3BP* has pleiotropic cellular functions, including roles in cellular migration, angiogenesis and immune system functioning (395). The gene has been implicated in the development of several forms of cancer (395). The gene product, LGALS3B*,* has recently been shown to be differentially expressed in non-alcoholic fatty liver disease (NAFLD), which is commonly associated with high BMI (396). Therefore, it is possible that *LGALS3BP* is associated with BMI because of its role in inflammation/immune system regulation, or because of its role in NAFLD.

6.9.1 Strengths and Limitations

The multiplex bisulfite-PCR-sequencing assay used in this study provided a high-throughput, cost-effective method of quantifying targeted CpG methylation at a single base resolution (383). However, it also presented methodological limitations.

Most evidently, there was a relatively low rate (79/192) of primer regions with adequate coverage for downstream analyses and a relatively low mapping efficiency of approximately 20%. The multiplexing of four primer pairs in each well of the Fluidigm Access Array likely contributed to these inefficiencies by increasing the rate of primer dimer formation. Other causes include different annealing temperatures of the primers, and sequence-dependent differences in PCR efficiency, i.e. that genetic variation in the primer sequence could have resulted in differences in the efficiency of PCR products (397). Mapping efficiency is generally decreased when mapping bisulfite treated sequences to a reference genome due to the reduced sequence complexity (386). Previous research has, however, suggested that a mapping efficiency of approximately 50-70% would have been expected from this type of experiment (398, 399). Not multiplexing the primers would have made the method

less cost-effective. However, in future experiments the balancing of cost versus a potentially improved rate of regions with enough coverage should be carefully considered. It is also striking that while no single previously identified obesity-associated CpG site replicated in this study, effect sizes in the Bis-PCR-Seq assay were significantly correlated to effect sizes of the original obesity EWASs.

6.9.2 Future Directions

The results of this study will be taken into account when refining the methodology of the large scale study of 1000 lean and 1000 obese males, such as in deciding whether or not to multiplex primers.

It would also be of value to identify sites associated with obesity in more disease relevant tissues, such as adipose or hepatic tissue. Further, coupling the study with functional gene output, such as transcriptome, analyses would yield insights into the interactions between epigenetic signatures of obesity and their physiological consequence.

The mortality associated with obesity does not come from having a high BMI, but from associated conditions such as cardiovascular events, some forms of cancer and chronic kidney disease (13). Therefore, it is of particular value to identify epigenetic biomarkers that are associated with an increased risk of developing such complications before they occur. This way, individuals could be offered targeted intervention or additional monitoring to reduce obesity associated mortality. In this regard, large-scale, longitudinal analyses with matched samples from individuals who gain weight would be of considerable value.

6.10 Summary

In this study, I aimed to validate previously identified obesity-associated CpG sites in a novel cohort of 96 obese men compared with 96 lean men. This was performed by using a highly multiplexed bisulfite-PCR-sequencing microfluidics-based assay. No CpG sites met the assay-wide threshold for an association with BMI. The study identified strengths and weaknesses with the bisulfite-PCR-sequencing methodology

used for validating CpG methylation associated with obesity. These insights will inform a large-scale project aiming to generate a robust, replicable obesity associated DNA methylation profile in a cohort of 1000 lean and 1000 obese males. The study also brought forward two CpG sites as being suggestive of an association with BMI that are replicable across different cohorts and study settings.

Elucidation of the epigenetic correlates of obesity has important implications for understanding disease pathogenesis and identifying biomarkers for disease progression. In a longer perspective, a more detailed understanding of how BMI-related health complications develop may aid the development of targeted therapeutic interventions that reduce the morbidity and mortality associated with obesity.

Chapter 7
General Discussion and Future Directions

7.1 Summary of Key Findings and Conclusions

I investigated the association between paternal metabolic health and offspring birth weight in a prospective cohort study of 500 mother-father-offspring trios (Chapter 3). Three groups of fathers (lean, overweight and obese) were recruited and these had clear differences in metabolic parameters including insulin resistance, waist circumference, blood pressure and lipid profiles. I collected detailed health data on their partner (the mother). I followed up these couples with regards to pregnancy outcome, particularly focussing on the birth weight of their offspring. Contrary to previous retrospective studies, I did not identify a paternal metabolic risk factor that significantly increased the risk of fathering small for gestational offspring (68, 69, 224-226). I did discover a non-significant trend towards higher insulin resistance in fathers of SGA infants. Due to limited numbers of SGA offspring and insulin-resistant fathers, I may have been under-powered to discover an association between paternal insulin resistance and growth restriction in his offspring. Such an association would be in line with results from studies of paternal monogenic diabetes, which is linked to a considerably reduced birth weight of his offspring (72). I did, however, find that paternal (own) birthweight was associated with his offspring's birth weight, while maternal (own) birthweight was not. The association between paternal own birth weight and offspring birth weight has been recognised previously (67).

My findings suggest that paternal genetic factors that influenced his own growth *in utero*, rather than factors acquired in his lifetime, are more likely to influence the intrauterine growth of his offspring. Conversely, maternal genetic factors that influenced her own growth *in utero* appear to have little effect on the *in utero* growth of her offspring, but rather the intra-uterine environment influences fetal growth.

In a separate study, I performed a systematic review of studies analysing DNA methylation in human sperm (Chapter 4). I critically evaluated 124 articles relevant to the topic in accordance with PRISMA guidelines and objectively rated the quality of evidence for each publication (191). For each of the three main research methodologies (analyses of global DNA methylation, candidate gene analyses and genome-wide analyses), I summarised findings from studies where the quality of

evidence was rated as high. I concluded that the human sperm methylome has most frequently been studied in the context of sub- or infertility. In this, studies have typically assessed the association between fertility and promoter methylation of a small number of candidate genes, most commonly imprinted genes. Findings from such studies have been inconsistent. Indeed, I identified that there was little overlap between the findings from candidate gene analyses and the results from epigenome wide association studies (EWASs) of sub- and infertility. I also concluded that compared to readily available somatic tissues, there is limited evidence that the human sperm methylome can be dynamically remodelled in response to environmental influences.

The systematic review also formed the basis of generating recommendations for future research. These included to shift research focus from candidate gene analyses and studies of global DNA methylation to well-powered genome wide approaches, e.g. by using microarrays (until whole-genome bisulfite sequencing becomes a cost-effective option). There is also a need to widen the scope from an overwhelming focus on fertility to more focus on the impact of metabolism, ageing and toxins on the sperm methylome. Future research should present detailed phenotyping of participants, in particular their age and smoking status. Such covariates should also be taken into account when analysing results, e.g. by using multivariate regression analyses. I also recommended that research should include more longitudinal analyses, for example examining how changes in environmental or physiological factors impact the sperm methylome over time. I recommended that for studies of tissue specificity of DNA methylation, samples should be matched (i.e. come from the same individual) to avoid confounding of results due to genetic variation. These and other conclusions from the systematic review informed the comprehensive characterisation of the human sperm methylome described in Chapter 5.

I performed the largest to date genome-wide characterisation of matched sperm and blood samples using the latest generation DNA methylation profiling array, the MethylationEPIC array (199) (Chapter 5). DNA methylation levels at > 850,000 CpG sites were measured in matched sperm and blood from a total of 92 individuals. I

found that, overall, sperm exhibited a highly polarised methylation profile towards the two extremes of DNA methylation levels, i.e. that both low (<20%) and high (>80%) levels of methylation were more commonly seen in sperm than in blood.

I specifically sought to identify CpG sites where sperm and blood methylations co-vary. This would allow blood to act as a proxy tissue in studies where sperm is unavailable. Significant correlation between sperm and blood methylation levels was identified at ~5,000 CpG sites. A high proportion of these sites displayed bi- and trimodal patterns of methylation (suggestive of a genetically driven effect) and that a high proportion had SNPs in the CpG site, as identified by cross-checking with the dbSNP Human Build 151 database (374). In other words, it is likely that methylation levels at the vast majority of these ~5,000 CpG sites was genetically driven. Indeed, there were only 10 CpG sites with significant methylation co-variation between the tissues and no known SNP in the probe sequence. These 10 CpG sites are likely to be too few to be of value in studies where blood is to be used as a surrogate tissue for sperm DNA methylation. It is, however, important to bear in mind that even at these 10 CpG sites where no SNP included in the dbSNP database was identified as present at the CpG site itself, methylation levels at these sites could still be influenced by genetic polymorphisms. For example, local cis-regulatory elements and genetic effects on the methylome via transcription factors cannot be ruled out as influencing CpG methylation at these 10 sites (382). Indeed, we are likely to currently be underestimating the genetic influences on DNA methylation when interpreting the results of EWASs (400). It is probable that results from present-day EWASs assumed to reflect epigenetic alterations in response to various phenotypic changes will be revised with an improved understanding of the influence of genetic variation on DNA methylation, for example by continued identification of methylation quantitative trait loci (mQTLs) (400).

I concluded that the DNA methylation profiles of sperm and blood are clearly distinct, driven by highly significant mean differences at multiple sites across the genome. There is little evidence of methylation co-variation between blood and sperm.

A comparison between methylation profiles of sperm and >6,000 somatic tissue samples available on the Gene Expression Omnibus was also performed. This analysis suggested that transcriptional regulation in spermatozoa is highly distinct from that of somatic tissue, most likely to repress overall transcription (265). My findings emphasise the importance of using disease-relevant tissues for epigenomic analyses, and question the validity of previous studies where blood has been used as a proxy tissue for sperm DNA methylation (153).

In the above study, I found no evidence that obesity alters the methylation covariation between sperm and blood. I did, however, identify one CpG site that was differentially methylated in sperm and blood of overweight/obese men compared to lean men. It is interesting to note that this CpG site has not previously been identified as associated with obesity/BMI in EWASs using blood only.

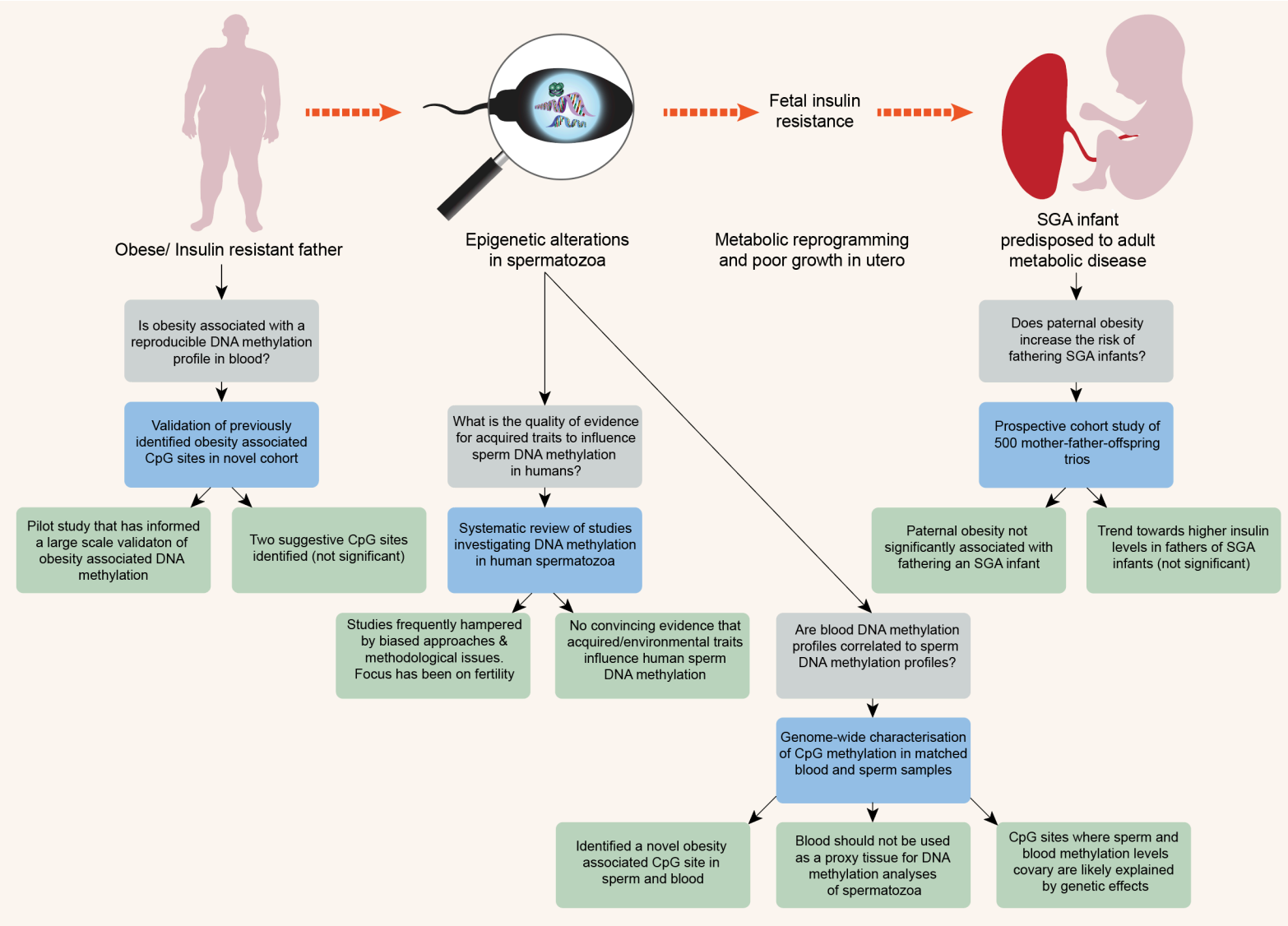
I validated previously identified obesity-associated CpG sites in a blood from a novel cohort of 96 obese and 96 lean males using a targeted bisulfite-PCR-sequencing approach (Chapter 6). This study constituted the first step in a large scale analysis aimed to identify a robust and replicable obesity associated DNA methylation profile. The larger study will include 1000 lean and 1000 obese males from the Norwegian Mother and Child cohort (MoBa) (401). Whilst I did not identify any CpG sites that were statistically associated with obesity at a genome-wide level, two CpG sites almost reached significance (FDR corrected p = 0.07). None of these CpG sites mapped to the 5' promoter of their nearest gene, and any said mechanistic involvement in the pathogenesis of obesity ascribed to these sites should be investigated with functional or other analyses. These and other important caveats in the interpretation of these EWAS results are discussed further in section 7.4. Whilst being cautious to ascertain whether cg07037944 actually influences the gene to which it is nearest, this CpG site is annotated to the gene Death Associated Protein Kinase 2 (*DAPK2*) (104). This CpG site has been previously been associated with BMI in at least two large-scale EWASs of obesity (104, 106). The *DAPK2* gene product has a role in mediating inflammatory reactions, and may therefore play a role in the chronic low-grade inflammation that is a feature of obesity (393). The second CpG

site, cg26651978, is annotated to the gene Lectin Galactoside-Binding Soluble 3-binding Protein (*LGALS3BP*) (104). Several CpG sites annotated to *LGALS3BP* have previously been identified as associated with BMI (102, 104, 107). *LGALS3BP* has also been found to be differentially expressed in non-alcoholic fatty liver disease (NAFLD), which is commonly associated with high BMI (396).

Importantly, this study also allowed evaluation of the bisulfite-PCR-sequencing method which will be used in the large-scale analysis. It identified strengths and limitations of the approach that will allow methodological refinement before the large study gets underway. This will improve our ability to generate a robust and replicable obesity associated DNA methylation profile, which has important implications for understanding disease pathogenesis, identifying biomarkers for disease, and identifying therapeutic targets.

An overview of research questions, approaches to investigations, key findings and conclusions of my PhD project in presented in Figure 7-1.

Figure 7-1. *The following page*: Summary of key findings in relation to the main hypotheses of my PhD project
A summary of the main hypotheses guiding this work is presented, along with key research questions (grey boxes), methods of investigation (blue boxes) and key findings/conclusions from my research (green boxes).
SGA = Small for Gestational Age


Fetal insulin resistance
Obese/ Insulin resistant father
Epigenetic alterations in spermatozoa
Metabolic reprogramming and poor growth in utero
SGA infant predisposed to adult metabolic disease
Is obesity associated with a reproducible DNA methylation profile in blood?
Validation of previously identified obesity associated CpG sites in novel cohort
Pilot study that has informed a large scale validaton of obesity associated DNA methylation
Two suggestive CpG sites identified (not significant)
What is the quality of evidence for acquired traits to influence sperm DNA methylation in humans?
Systematic review of studies investigating DNA methylation in human spermatozoa
Studies frequently hampered by biased approaches & methodological issues. Focus has been on fertility
No convincing evidence that acquired/environmental traits influence human sperm DNA methylation
Are blood DNA methylation profiles correlated to sperm DNA methylation profiles?
Genome-wide characterisation of CpG methylation in matched blood and sperm samples
Identified a novel obesity associated CpG site in sperm and blood
Blood should not be used as a proxy tissue for DNA methylation analyses of spermatozoa
CpG sites where sperm and blood methylation levels covary are likely explained by genetic effects
Does paternal obesity increase the risk of fathering SGA infants?
Prospective cohort study of 500 mother-father-offspring trios
Paternal obesity not significantly associated with fathering an SGA infant
Trend towards higher insulin levels in fathers of SGA infants (not significant)

7.2 Implications of my Research Findings

Over the course of my PhD, I employed a range of methods to interrogate whether, and how, paternal metabolic health influences sperm DNA methylation and the intrauterine growth of his offspring. In doing so, this thesis has contributed to the field of intergenerational inheritance and obesity associated epigenetic change in humans in a number of ways.

First, this project has demonstrated that paternal BMI is associated with only a modest, if any, effect on offspring birth weight. A larger study or the study of extreme phenotypes would be necessary to identify an association (68, 69). It could be seen as encouraging that my research and a recent systematic review did not find evidence for paternal obesity at conception as associated with offspring birth weight (227). Instead, a meta-analysis of the association between paternal BMI at conception and offspring health outcomes showed that high paternal BMI around conception increased the risk of his offspring having a high BMI or high fat mass in childhood (227). This association could have arisen from factors in paternal spermatozoa that increase the risk of high adiposity in childhood, from lifestyle factors (an 'obesogenic environment'), or from a combination of the two. In contrast to factors transferred via sperm during conception, an obesogenic environment is modifiable after conception has occurred. Further, pregnancy represents a period of increased motivation for parents to undertake lifestyle changes, and of coming into contact with health care providers (402). If it is the case that paternal BMI at conception is less influential than paternal BMI during a child's early years, then intervening by encouraging healthy lifestyle habits (for both parents) during pregnancy could have considerable benefits for future offspring health.

In line with previous research, I showed that paternal (own) birth weight is associated with the birth weight of his offspring (67). Antenatal prediction of estimated infant birth weight is important, as it guides both the timing and mode of delivery of a pregnancy (403). Although ultrasound based estimation of fetal weight has improved considerably in the last decades, it can still be inaccurate, especially if performed in late gestation (403). In a longer perspective, therefore, one could envisage that

paternal (own) birth weight is also taken into account to achieve a more accurate prediction of fetal weight.

This thesis includes the first systematic review of studies of human sperm DNA methylation. The detailed overview of the field generated via the systematic review should be a useful resource for other researchers in the area, and for those interested in epigenetics of human germ cells. In this, I summarised findings from high-quality studies of global DNA methylation, candidate gene studies and genome-wide investigations of spermatozoal DNA methylation respectively. I found that few, if any, of the findings from candidate gene analyses of, typically, fertility correlated with findings from less biased genome-wide studies of the same phenotype. I also concluded that there are few, if any, robust and reproducible epigenetic correlates of subfertility from EWASs performed thus far. In this regard, it is interesting that there is already work underway to implement epigenetic testing of sperm as part of clinical semen analyses (404). It is possible that future, large-scale, DNA methylation analyses of spermatozoa will indeed identify specific CpG sites that can serve as reliable biomarkers of different types of subfertility that will be of clinical value. Until then, however, based on the inconsistent findings of studies of DNA methylation of fertility included in my systematic review, the clinical utility of designing a platform for measuring spermatozoal DNA methylation is likely to be limited. The critical evaluation of studies included in this systematic review allowed me to make specific and constructive recommendations for future research (summarised in section 7.1 above). I hope that these recommendations will contribute towards refining research methodologies in this field, ultimately leading to clinically meaningful results regarding how paternal environmental and acquired traits can influence the next generation and beyond.

Third, this project has empirically demonstrated the necessity of using disease relevant tissues, i.e. germ cells, in epigenomic investigations of intergenerational inheritance. Previous studies of intergenerational effects of metabolic phenotypes have used results from epigenetic analyses of blood and pancreatic tissue to infer transmission of specific DNA methylation signatures via spermatozoa (138, 153). My

findings strongly question the validity of this approach by showing little, if any, evidence of methylation co-variation between sperm and blood.

I have demonstrated that a targeted bisulfite-PCR-Sequencing (Bis-PCR-Seq) approach can be a high-throughput, cost-effective alternative to microarray based methods of interrogating CpG methylation in large cohorts, e.g. in the context of validating previously identified CpG sites. Although Bis-PCR-Seq had drawbacks in my study, for example in that a large proportion of CpG sites could not be analysed due to low read coverage, we showed a high overall correlation of results between this method and using the MethylationEPIC array. After further refinement of the approach, in particular by improving strategies to reduce primer dimerisation, we aim to use this approach to generate a robust obesity associated DNA methylation profile in a cohort of 1,000 lean and 1,000 obese males.

7.3 Strengths of my PhD Project

Most previous studies analysing the impact of paternal BMI and insulin resistance on offspring birth weight have been retrospective (68, 69, 225, 226). The prospective nature of the Dad's Health Study reduced potential selection bias and strengthened my ability to make results generalisable to a wider population. The Dad's Health Study also included a more detailed phenotypic assessment of male participants than has often previous been the case. For example, fasting blood tests allowed us to reliably assess paternal insulin resistance, which is likely to be a more accurate measure of metabolic health compared to BMI alone (68).

The analysis of genome-wide DNA methylation of matched sperm and blood samples (n = 92) had a considerably larger sample size compared to previous studies with a comparable methodology (n = 8) (168). In contrast to most previous investigations of spermatozoal DNA methylation, participants were not recruited from a reproductive medicine setting but were healthy, fertile volunteers. This makes findings from the analysis more generalisable to the male population as a whole. Another strength of this study was the use of the most recent DNA methylation profiling array, the MethylationEPIC array. This allowed characterisation of genomic regions that have

previously only been analysed in in human spermatozoa in small-scale studies (<10 participants) using whole-genome bisulfite sequencing (341, 352).

Another strength of the work presented in this thesis is that it includes novel methods of analysing epigenetic data. With co-workers, I used a recently developed data analysis pipeline designed for large methylation data sets to compare spermatozoal DNA methylation patterns to that of >6,000 somatic tissue samples (211). To my knowledge, this is the most extensive comparison between the DNA methylation profile of spermatozoa and that of somatic tissue samples.

7.4 Limitations of my PhD Project

I have discussed specific limitations of the studies described in each chapter separately. I will therefore limit the discussion below to more general and recurring limitations relevant to my PhD project.

I had limited sample sizes that particularly affected the Dad's Health Study described in Chapter 3. This likely made the study underpowered to detect modest effects of paternal influences on offspring birth weight. The power calculations used to determine a sample size for the Dad's Health Study were based on effect sizes in previous studies which might have included confounding factors that influenced fetal growth by other mechanisms than acquired paternal traits, such as genetic influences or residual confounding not addressed in customised birthweight centiles (68, 176). Also, whilst previous studies in this area have been able to detect effects with sample sizes comparable to or smaller than the Dad's Health Study cohort, they may have included more extreme phenotypes, e.g. severe growth restriction (69, 72). The Dad's Health study mostly included healthy mother-father-offspring trios with most offspring appropriately grown for gestational age. Also, the Dad's Health Study included relatively few obese fathers compared with lean and overweight fathers. Increasing the number of obese fathers would improve power and possibly unmask an association between paternal insulin resistance and fetal growth restriction. Including fathers with diabetes might also strengthen the link between paternal insulin resistance and fetal growth restriction. I identified a trend towards higher

insulin resistance in fathers of low birth weight infants that could be confirmed with a larger study size. Alternatively, studying offspring born to men with type 1 diabetes who have no insulin compared with men who have type 2 diabetes with generally high insulin levels could be revealing.

Another potential source of imprecision applicable to the Dad's Health study is lack of standardisation for weighing infants. For the Dad's Health study, the birth weight of infants was obtained from UCLH discharge summaries. It is possible that some infants were less dry than others when the weighing was performed, and that there could have been differences in the time elapsed since the birth of an infant and the infant being weighed. It would have been prudent to standardise weighing of infants included in the study, for example weighing all infants at 30-45 minutes after birth and taking care that infants were dry when weighed.

Additionally, in the context of sample sizes, the cohort of obese males (n = 24) who provided matched samples of blood and sperm was likely too limited to detect modest effects on how obesity may influence the covariation of DNA methylation between blood and sperm. The fact that I nevertheless identified one novel obesity associated CpG site in blood and sperm points towards the possibility that there is more to be explored in terms of gamete-specific epigenetic correlates of obesity that may not be detected in studies that only analyse blood.

A further limitation is that while several strategies were used to minimise confounding of results in the analysis of how paternal metabolic health influences fetal growth, it is still possible that there were residual confounders not taken into account. One of these is socioeconomic status. There is a clear and consistent association between low socioeconomic status and low birth weight, even across more economically developed countries with state-funded healthcare systems such as the UK (405-407). Low socioeconomic status is also strongly linked to an increased risk of obesity and T2DM (408-410). In other words, low socioeconomic status increases the risk both of metabolic disease and of fathering low birth weight offspring. This could yield an association between paternal insulin resistance and low

birth weight offspring, without insulin resistance necessarily being the causal mechanism. It is possible that this factor may have influenced previous retrospective and case-control studies that identified an association between paternal obesity or insulin resistance and low offspring birth weight (69). In either case, it is a limitation of the Dad's Health study that detailed information on household income and other determinants of socioeconomic status were not collected and taken into account.

Another limitation of this project is the focus on DNA methylation as opposed to other epigenetic mechanisms and their integration. Early studies of intergenerational inheritance of acquired traits suggested that DNA methylation was a likely driver, or at least an important component of, mediating these effects (138, 140, 141). However, during the time of working on this project, the focus of analyses of intergenerational inheritance has increasingly shifted towards small RNA species (144, 145). Indeed, one study identified specific small RNA molecules capable of mediating effects of metabolic disease between generations (145). Epigenetic regulation by small RNA species and DNA methylation is of course not mutually exclusive, and there is extensive interaction between these mechanisms (411). Nevertheless, limiting my research to DNA methylation may have narrowed the potential for identifying biologically meaningful findings (145).

Other limitations of the work presented in thesis relate to challenges in the interpretation of epigenome wide analyses of particular phenotypes more broadly. In this regard, one of the key issues is our yet incomplete understanding of, and methods for investigating, the association between sequence variation and DNA methylation. While estimates vary widely, it has been suggested that as much as 22% to 80% of the variability of DNA methylation van be accounted for by underlying genetic variation (412). In our studies, we accounted for genetic polymorphisms e.g. by removing probes from DNA methylation data using annotated lists of SNPs in the CpG site. In the study of methylation covariation between sperm and blood we further cross-checked results against the dbSNP database to investigate whether SNPs at the particular CpG sites investigated were likely to influence results (374). However, although these methods are among the currently most commonly used

methods to account for the influence of sequence variation on DNA methylation data, they are unlikely to be sufficient (412, 413). Indeed, including matched genotype information on participants along with information on methylation quantitative trait loci (mQTLs) will likely be required to achieve a more reliable understanding of whether phenotypic outcomes are associated with particular epigenetic signatures.

It should be appreciated that regions of the genome that are involved in regulating the expression of a particular gene can be far away from the gene itself, for example in an enhancer (413). In this regard, the results from the validation study of obesity associated DNA methylation in blood should be interpreted with caution. Thus, none of the two CpG sites that were suggestive of being reproducibly associated with obesity (Chapter 6) were in the 5' promoters of their nearest gene, where they would have been most likely to directly influence their expression. One of them (cg26651978) was 3' of LGALS3BP, and could potentially be involved in regulating a distant gene, e.g. by being in an enhancer region. Therefore, although both DAPK2 and LGALS3BP have plausible roles in the pathogenesis of obesity, it is far from clear that they are involved in the disease phenotype based on current EWAS findings.

It is also worth reiterating that all large-scale EWASs of obesity and related traits thus far have been performed on the Illumina HM450. These include the studies from which we selected CpG sites for replication in the study of obesity associated DNA methylation in blood. Not only does this array interrogate less than 2% of the CpG sites in the human genome, but it is also heavily biased towards promoter regions, which may not be the most relevant sites for studies of phenotypic variation (414).

As previously discussed (section 1.8.4), a major challenge in the interpretation of findings from EWASs is determining causality, i.e. whether particular epigenetic marks occur as a cause or a consequence of a particular phenotype (412, 413). This is less of a concern if the aim of a study is to, for example, identify biomarkers of a disease. However, if the aim is to appreciate if and how epigenetic alterations can influence a particular disease process, then methods of determining the causal

relationship between DNA methylation and phenotypic change should be employed. Such methods include two-step Mendelian randomisation, which aims to ascertain the causal relationships between exposure, DNA methylation and outcome (415). Using this method on large scale EWASs of obesity performed thus far indicate that a majority of DNA methylation signatures associated with obesity are a consequence rather than a cause of the disease process (106). This should be borne in mind when, for example, interpreting findings from the validation study of obesity associated CpG methylation in blood; that the two suggestive findings are potentially the consequence rather than the cause of a high BMI. In future studies, longitudinal analyses of individuals recruited prior to the onset of a particular phenotypic change will also be key in avoiding issues around, for example, reverse causality in the context of epigenomic investigations (412, 414).

As a consequence of the difficulties in interpreting results from EWASs, such as those described above, several researchers have argued that no EWAS performed to date can be said to be fully interpretable (400, 412).

Lastly, the power calculations employed to determine an appropriate sample size for the Dad's Health study (section 2.3.4) were problematic not only because they only took into account two groups of lean and obese males respectively, rather than three groups of lean, overweight and obese males. Indeed, one of the approaches for determining an appropriate sample size was based on a power of only 80% (rather than the more stringent 90% which was employed in the other approach). This may have limited our ability to identify clinically meaningful influences of paternal metabolic health on fetal growth.

7.5 Future Directions

As frequently is the case, the work presented in this thesis has opened up for several additional lines of enquiry, some of which are already underway.

7.5.1 Investigating the potential for reversibility of obesity associated epigenetic markers

In the context of public health policy implications, it is of particular interest to explore whether obesity associated epigenetic signatures are reversible. If so, improvements in a man's health before conception would make him less likely to impact on his offspring via epigenetic alterations passed via his sperm. I have therefore commenced recruitment for a prospective cohort study of obese males (BMI ≥ 35 kg/m^2) due to undergo bariatric surgery. The study analyses serial blood and semen samples from 15 obese men due to undergo bariatric surgery. In this, one set of blood and semen samples is collected approximately one month prior to surgery and constitutes the baseline, and a second set is collected three to four months after the surgery (when one full round of spermatogenesis has taken place). The third and final set is collected 9-12 months following bariatric surgery, when most of the weight loss has occurred (416). The study will also include a control cohort of 15 obese males not undergoing weight loss surgery, but providing samples at comparable times. This will help to reduce confounding by factors such as ageing. Such a control group has been lacking in previous research of weight loss related changes in spermatozoa (346).

A previous small-scale study of six obese males undergoing bariatric surgery indicated that bariatric surgery could be associated with widespread epigenetic alterations in male germ cells. However, possibly due to the limited sample size, the study failed to identify any CpG sites that fulfilled the threshold for genome wide significance (147). Further, the study did not include a control cohort interrogated for spermatozoal DNA methylation changes occurring due to processes different from the bariatric surgery, such as ageing.

Results from the bariatric surgery study will be a powerful way to inform public health policies directed towards improving preconception health. If obesity-associated epigenetic signatures present of spermatozoa are reversible with weight loss, then targeted intervention strategies that optimise a man's health prior to conception have the potential to improve the health of the next generation and beyond.

7.5.2 Expanding sample sizes to detect modest effects of paternal insulin resistance on fetal growth

The Dad's Health study was likely underpowered to detect modest effects of paternal metabolic traits on offspring with birth weights across the normal birth weight spectrum. However, I did identify a non-significant trend of higher insulin resistance in fathers of low birth weight offspring, in spite of there being no similar trend in paternal BMI. This suggests the possibility that with an increased sample size, paternal insulin resistance might emerge as significantly associated with fathering low birth weight offspring. This association has clearly been demonstrated in the case of paternal monogenic diabetes (72). Several observational studies have also suggested an association between paternal insulin resistance and low offspring birth weight (69, 224, 225). To my knowledge, however, this has never been demonstrated in a prospective study. I therefore aim to expand the Dad's Health study to further interrogate whether insulin resistance, rather than the more crude measure of paternal BMI, may be a modifiable risk factor for fathering low birth weight infants.

7.5.3 Investigating longer-term effects of paternal metabolic disease

Future work should also be focussed on longer term health consequences of paternal metabolic disease. It is possible that offspring born to males with metabolic disease may not display overt growth restriction or metabolic problems at birth, but instead do so later in life. For example, previous epidemiological studies have indicated that offspring born with a low weight at birth are particularly vulnerable to detrimental health consequences if they undergo rapid 'catch-up growth' in their first few years of life (66). Therefore, following up children born to fathers with metabolic disease by measuring e.g. BMI, waist circumference and insulin resistance could yield insights into intergenerational effects of metabolic disease that may not be obvious at birth. The challenge with this type of study in humans would be to account for the influences of a shared 'obesogenic' environment. Nevertheless, such research could inform targeted public health policies directed towards families particularly at risk of metabolic disease.

7.5.4 Other avenues of investigating intergenerational effects in humans

One could argue that paternal smoking and aging have more robust and replicable effects on offspring health than does paternal obesity (227). Paternal smoking has more consistently than paternal obesity been linked to low offspring birth weight (69, 227). High paternal age at conception has consistently been linked to an increased risk of some psychiatric disorders in his offspring (227, 417). Both age and smoking are reliably known to influence DNA methylation signatures (195, 196). Age also appears to predictably influence DNA methylation signatures of spermatozoa (372). Therefore, investigating germline epigenetic signatures of smoking and aging could yield more consistent findings in terms of their potential to influence the next generation. This may be an interesting avenue for future research that could also become a proof of principle for intergenerational effects of epigenetic change in humans.

7.5.5 Continued interrogation of spermatozoal regulation of gene expression

The comprehensive characterisation of genome-wide CpG methylation in spermatozoa described in Chapter 5 identified a number of specific CpG sites that should be taken forward for continued investigation. For example, there were sites that showed a uniform level of methylation in one of the tissues, but a bimodal or trimodal pattern of methylation in the other. This begs the question of how tissue specific transcription factors regulate leukocyte- and spermatozoal gene expression at these sites. Also, we identified CpG sites with a negative correlation of methylation between sperm and blood, i.e. that higher methylation in blood is correlated to lower methylation in sperm. This suggests that there are physiological or environmental factors that increase methylation levels in one tissue but decreases them in the other. Characterisation of such factors could yield fascinating insights into tissue-specific, dynamic regulation of gene expression.

Following publication of our findings of DNA methylation in matched sperm and blood samples, we will make our data publicly available in manner equivalent to what has been done for studies of e.g. methylation correlation of blood and brain tissue

(368). This should be a valuable resource for the interpretation of data from blood-based EWAS analyses.

It will also be interesting to see if the CpG site that we identified as significantly associated with obesity in sperm and blood replicates in other studies of obesity associated DNA methylation in sperm. If so, we may have identified a hint of an obesity associated epigenetic signature in a tissue that has the potential to influence the next generation.

7.5.6 Large scale analysis of obesity associated DNA methylation profiles in blood

The study described in Chapter 6, in which I validated previously identified obesity associated CpG sites in a novel cohort of 96 lean and 96 obese males, constituted the first step in a large-scale analysis aimed to identify a robust and replicable obesity associated DNA methylation profile. The larger study will include 1000 lean and 1000 obese males from the Norwegian Mother and Child cohort (MoBa) (401) (MRC reference code MR/P011799/1; title *'Paternal obesity-associated DNA methylation: an investigation into its reproducibility, reversibility and association with fetal growth restriction'*).

Results from this investigation will be of value in understanding the pathogenesis of obesity, developing biomarkers for disease and identifying therapeutic targets. In addition, we aim to use results from this study to investigate whether paternal obesity could influence his offspring in ways that were not obvious from the studying offspring birth weight alone. With this aim, I have collected cord blood samples from the majority of infants born to parents participating in the Dad's Health Study. The CpG sites that emerge as robustly associated with obesity in the large scale validation study will be analysed for enrichment in cord blood of infants born to obese fathers compared to cord blood from infants born to lean fathers. Some previous studies have indicated that paternal obesity is indeed associated with altered offspring DNA methylation profiles, however these studies have employed a candidate gene approach and been relatively small scale (153, 418).

7.5.7 Distinguishing between genetic and acquired paternal effects on fetal growth

A significant challenge in interpreting results from studies suggesting that paternal obesity or insulin resistance increases the risk of fathering low birth weight offspring has been to distinguish between what is an effect of paternal genetics versus acquired paternal traits. More specifically, two studies that informed my PhD project both showed that whilst obese or insulin resistant men were more likely to father low birth weight offspring, they were also more likely to have been small at birth themselves (68, 69).

Interestingly, novel research methods coupled with considerably increased sample sizes (>550,000 participants in total) have begun to distinguish between effects on fetal growth that are genetically driven by the mother versus those that are genetically driven by the fetus and those driven by an interaction between the two (419). In a longer perspective, it would be interesting to develop this research strategy further to tease apart the relative contributions of paternal genetic and paternal non-genetic effects on fetal growth. Such a study would of course also need to account for maternal genetic and non-genetic effects. As indicated by the above study, a sample size of hundreds of thousands of DNA samples from mother-father-offspring trios are likely to be required to achieve this aim.

7.6 Summary and Concluding Remarks

The concept of epigenetic markers as mediators of inter- and transgenerational effects of acquired traits is still controversial. Indeed, some researchers suggest that stochastic genetic and epigenetic variation have more influence on the sperm epigenome than, for example, dietary insults (143). The work presented in this thesis does not prove epigenetic change as underlying associations between acquired parental phenotypes and offspring health. It does, however, present a number of findings which advance the field of epigenetics in intergenerational inheritance.

I have performed the largest to date genome-wide characterisation of matched human sperm and blood samples. From this, I have demonstrated that the highly

discordant and practically completely uncorrelated DNA methylation profiles of sperm and blood necessitates studying germ cells, rather than proxy tissues, in analyses of intergenerational effects. Such findings are in accordance with previous research that has analysed the methylation co-variation between blood and less readily available tissues such as brain (368). I have also identified specific CpG sites in spermatozoa that are of particular interest in understanding sperm-specific regulation of gene expression, and the factors that it may be influenced by. Further, I have identified a novel obesity associated CpG site in sperm and blood that has not previously been identified in EWASs of blood only. Interestingly, however, methylation at the same CpG site has previously been identified as associated with obesity in liver (377).

Comparison of spermatozoal DNA methylation to the DNA methylation profiles of >6,000 somatic tissue samples available on the Gene Expression Omnibus database allowed us to confirm that spermatozoal transcriptional regulation is highly distinct, most likely due cause overall transcriptional repression (210, 265).

Further, I have performed a comprehensive systematic review of studies of DNA methylation in human spermatozoa, and generated a number of concrete recommendations for future research based on my findings.

In a prospective cohort study, I have shown that the impact of paternal metabolic health on offspring birth weight is limited, at least across the spectrum of normal birth weights. From this, I have suggested ways forward in further elucidating the role of paternal traits, in particular insulin resistance, on the fetal growth of his offspring. Continuing this work is of considerable importance from a public health perspective as fetal development represents a crucial time in which interventions could have enduring and powerful effects on adult health and disease.

I have begun to validate previously identified CpG sites associated with obesity with the aim to generate a robust and reproducible DNA methylation signature of obesity in blood.

Human obesity and birth weight are incredibly complex phenotypes influenced by a multitude of genetic, environmental, gene-environment, behavioural and social factors. Research in this area is subject to numerous challenges. There is, however, enormous potential gains from their continued elucidation, both scientifically and from a public health perspective. Research presented in this thesis constitutes an important step towards this aim.

References

1. Sturm R, An R, Maroba J, Patel D. The effects of obesity, smoking, and excessive alcohol intake on healthcare expenditure in a comprehensive medical scheme. Samj South African Medical Journal. 2013;103(11):840-4.
2. World Health Organization. WHO Fact Sheet: Obesity and overweight Geneva: WHO; 16 February 2018 [Available from: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.
3. Scarborough P, Bhatnagar P, Wickramasinghe KK, Allender S, Foster C, Rayner M. The economic burden of ill health due to diet, physical inactivity, smoking, alcohol and obesity in the UK: an update to 2006-07 NHS costs. Journal of Public Health. 2011;33(4):527-35.
4. Di Cesare M, Bentham J, Stevens GA, Zhou B, Danaei G, Lu Y, et al. Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19.2 million participants. Lancet. 2016;387(10026):1377-96.
5. Ford ND, Patel SA, Narayan MV. Obesity in Low- and Middle-Income Countries: Burden, Drivers, and Emerging Challenges. Annual Review of Public Health, Vol 38. 2017;38:145-64.
6. Baker C. House of Commons Library Briefing Paper Number 3336: Obesity Statistics. United Kingdom: House of Commons; 2018.
7. Public Health England. Diabetes prevalence estimates for local populations United Kingdom: gov.uk; 2015 [Available from: https://www.gov.uk/government/publications/diabetes-prevalence-estimates-for-local-populations.
8. Moody A, Cowley G, Fat LN. Social inequalities in prevalence of diagnosed and undiagnosed diabetes and impaired glucose regulation in participants in the Health Surveys for England series (vol 6, e010155, 2016). Bmj Open. 2016;6(4).
9. Kanavos P, van den Aardweg S, Schurer W. Diabetes expenditure, burden of disease and management in 5 EU countries. United Kingdom: London School of Economics; 2012.
10. UK Government. Cm 9467: Public Expenditure Statistical Analyses 2017 United Kingdom: Chief Secretary to the Treasury; 2017 [Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/630570/60243_PESA_Accessible.pdf.

11. Public Health England. Health Survey for England - Health, social care and lifestyles United Kingdom: NHS England; 1993-2017 [cited 2019. Available from: https://digital.nhs.uk/data-and-information/areas-of-interest/public-health/health-survey-for-england-health-social-care-and-lifestyles.
12. International Diabetes Federation. The IDF consensus worldwide definition of the metabolic syndrome International Diabetes Federation; 2006. Available from https://www.idf.org/e-library/consensus-statements/60-idfconsensus-worldwide-definitionof-the-metabolic-syndrome.html.
13. Afshin A, Forouzanfar MH, Reitsma MB, Sur P, Estep K, Lee A, et al. Health Effects of Overweight and Obesity in 195 Countries over 25 Years. New England Journal of Medicine. 2017;377(1):13-27.
14. Pi-Sunyer X. The Medical Risks of Obesity. Postgraduate Medicine. 2009;121(6):21-33.
15. Stubert J, Reister F, Hartmann S, Janni W. The Risks Associated With Obesity in Pregnancy. Deutsches Arzteblatt International. 2018;115(16):276-+.
16. Ng M, Fleming T, Robinson M, Thomson B, Graetz N, Margono C, et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2014;384(9945):766-81.
17. Fildes A, Charlton J, Rudisill C, Littlejohns P, Prevost AT, Gulliford MC. Probability of an Obese Person Attaining Normal Body Weight: Cohort Study Using Electronic Health Records. American Journal of Public Health. 2015;105(9):E54-E9.
18. Whitaker KL, Jarvis MJ, Beeken RJ, Boniface D, Wardle J. Comparing maternal and paternal intergenerational transmission of obesity risk in a large population-based sample. American Journal of Clinical Nutrition. 2010;91(6):1560-7.
19. Lifshitz F. **Obesity in children Review**. J Clin Res Pediatr Endocrinol. 2008.
20. Department of Health. Healthy Lives, Healthy People: A call to action on obesity in England 2011 [Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/213720/dh_130487.pdf.
21. Willemsen G, Ward KJ, Bell CG, Christensen K, Bowden J, Dalgard C, et al. The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers: The Discordant Twin

(DISCOTWIN) Consortium. Twin Research and Human Genetics. 2015;18(6):762-71.
22. Herrera BM, Lindgren CM. The Genetics of Obesity. Current Diabetes Reports. 2010;10(6):498-505.
23. Wainschtein P, DPYengo, L Zheng, Z TOPMed Anthropometry Working GroupTrans-Omics for Precision Medicine Consortium Cupples , L Shadyab, A McKnight , B Shoemaker , B Mitchell , B Psaty, BKooperberg, CRoden, DDawood, DArnett, DRegan, EBoerwinkle, ERotter, JAllison, MMcDonald, M-LChung, MSmith, NEllinor, PVasan, RMathias, RRich, SHeckbert, SRedline, SGuo, XChen, ILiu, C-Tde Andrade, M,Yanek, L,Albert, CHernandez, RMcGarvey, SNorth, KLange, LWeir, BLaurie, CYang, JVisscher, P. Recovery of trait heritability from whole genome sequence data. bioRxiv. 2019.
24. Mayhew AJ, Meyre D. Assessing the Heritability of Complex Traits in Humans: Methodological Challenges and Opportunities. Current Genomics. 2017;18(4):332-40.
25. Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. Nature Reviews Genetics. 2013;14(2):139-49.
26. Prasad RB, Groop L. Genetics of Type 2 Diabetes-Pitfalls and Possibilities. Genes. 2015;6(1):87-123.
27. Singh RK, Kumar P, Mahalingam K. Molecular genetics of human obesity: A comprehensive review. Comptes Rendus Biologies. 2017;340(2):87-108.
28. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.
29. Fairbrother U, Kidd E, Malagamuwa T, Walley A. Genetics of Severe Obesity. Current Diabetes Reports. 2018;18(10).
30. Farr OM, Gavrieli A, Mantzoros CS. Leptin applications in 2015: what have we learned about leptin and obesity? Current Opinion in Endocrinology Diabetes and Obesity. 2015;22(5):353-9.
31. Kuhnen P, Krude H, Biebermann H. Melanocortin-4 Receptor Signalling: Importance for Weight Regulation and Obesity Treatment. Trends in Molecular Medicine. 2019;25(2):136-48.
32. Yeo GSH, Lank EJ, Farooqi IS, Keogh J, Challis BG, O'Rahilly S. Mutations in the human melanocortin-4 receptor gene associated with severe familial obesity disrupts receptor function through multiple molecular mechanisms. Human Molecular Genetics. 2003;12(5):561-74.

33. Huszar D, Lynch CA, FairchildHuntress V, Dunmore JH, Fang Q, Berkemeier LR, et al. Targeted disruption of the melanocortin-4 receptor results in obesity in mice. Cell. 1997;88(1):131-41.
34. Yeo GSH. Genetics of obesity: can an old dog teach us new tricks? Diabetologia. 2017;60(5):778-83.
35. Goodarzi MO. Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications. Lancet Diabetes & Endocrinology. 2018;6(3):223-36.
36. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature Genetics. 2010;42(11):937-U53.
37. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Felix R, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197-U401.
38. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nature Genetics. 2015;47(10):1114-+.
39. Polychronakos C, Alriyami M. Diabetes in the post-GWAS era. Nature Genetics. 2015;47(12):1373-4.
40. Misra S, Owen KR. Genetics of Monogenic Diabetes: Present Clinical Challenges. Current Diabetes Reports. 2018;18(12).
41. McDonald TJ, Ellard S. Maturity onset diabetes of the young: identification and diagnosis. Annals of Clinical Biochemistry. 2013;50(5):403-15.
42. Naylor R, Knight Johnson A, del Gaudio D. Maturity-Onset Diabetes of the Young Overview In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle2018 May 24.
43. Lau HH, Ng NHJ, Loo LSW, Jasmen JB, Teo AKK. The molecular functions of hepatocyte nuclear factors - In and beyond the liver. Journal of Hepatology. 2018;68(5):1033-48.
44. Anik A, Catli G, Abaci A, Bober E. Maturity-onset diabetes of the young (MODY): an update. Journal of Pediatric Endocrinology & Metabolism. 2015;28(3-4):251-63.
45. Dukes ID, Sreenan S, Roe MW, Levisetti M, Zhou YP, Ostrega D, et al. Defective pancreatic beta-cell glycolytic signaling in hepatocyte nuclear factor-1 alpha-deficient mice. Journal of Biological Chemistry. 1998;273(38):24457-64.

46. Verhave JC, Bech AP, Wetzels JFM, Nijenhuis T. Hepatocyte Nuclear Factor 1 beta-Associated Kidney Disease: More than Renal Cysts and Diabetes. Journal of the American Society of Nephrology. 2016;27(2):345-53.
47. Scott RA, Scott LJ, Maegi R, Marullo L, Gaulton KJ, Kaakinen M, et al. An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes. 2017;66(11):2888-902.
48. Kooner JS, Saleheen D, Sim X, Sehmi J, Zhang WH, Frossard P, et al. Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. Nature Genetics. 2011;43(10):984-U94.
49. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature Genetics. 2012;44(9):981-+.
50. Ma RCW, Hu C, Tam CH, Zhang R, Kwan P, Leung TF, et al. Genome-wide association study in a Chinese population identifies a susceptibility locus for type 2 diabetes at 7q32 near PAX4. Diabetologia. 2013;56(6):1291-305.
51. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nature Genetics. 2018;50(11):1505-+.
52. Hales CN, Barker DJP, Clark PMS, Cox LJ, Fall C, Osmond C, et al. Fetal and Infant Growth and Impaired Glucose Tolerance at age 64. British Medical Journal. 1991;303(6809):1019-22.
53. Yu ZB, Han SP, Zhu GZ, Zhu C, Wang XJ, Cao XG, et al. Birth weight and subsequent risk of obesity: a systematic review and meta-analysis. Obesity Reviews. 2011;12(7):525-42.
54. Poston L, Caleyachetty R, Cnattingius S, Corvalán C, Uauy R, Herring S, et al. Preconceptional and maternal obesity: epidemiology and health consequences. The Lancet Diabetes and Endocrinology. 2016:1025–36.
55. Kampmann U, Madsen LR, Skajaa GO, Iversen DS, Moeller N, Ovesen P. Gestational diabetes: A clinical update. World Journal of Diabetes. 2015;6(8):1065-72.
56. Webber J, Charlton M, Johns N. Diabetes in pregnancy: management of diabetes and its complications from preconception to the postnatal period (NG3). British Journal of Diabetes and Vascular Disease. 2015;15(3):107-11.

57. Rani PR, Begum J. Screening and Diagnosis of Gestational Diabetes Mellitus, Where Do We Stand. Journal of Clinical and Diagnostic Research. 2016;10(4):QE1-QE4.
58. Catalano PM, Mouzon SHD. Is it time to revisit the Pedersen hypothesis in the face of the obesity epidemic? American Journal of Obstetrics and Gynecology. 2011;204(6):479-87.
59. Pedersen J, Bojsenmoller B, Poulsen H. BLOOD SUGAR IN NEWBORN INFANTS OF DIABETIC MOTHERS. Acta Endocrinologica. 1954;15(1):33-52.
60. Hales CN, Barker DJP. The thrifty phenotype hypothesis. British Medical Bulletin. 2001;60:5-20.
61. Hanson MA, Gluckman PD. EARLY DEVELOPMENTAL CONDITIONING OF LATER HEALTH AND DISEASE: PHYSIOLOGY OR PATHOPHYSIOLOGY? Physiological Reviews. 2014;94(4):1027-76.
62. Roseboom TJ, van der Meulen JHP, Ravelli ACJ, Osmond C, Barker DJP, Bleker OP. Effects of prenatal exposure to the Dutch famine on adult disease in later life: an overview. Molecular and Cellular Endocrinology. 2001;185(1-2):93-8.
63. Knop MR, Geng TT, Gorny AW, Ding RY, Li CW, Ley SH, et al. Birth Weight and Risk of Type 2 Diabetes Mellitus, Cardiovascular Disease, and Hypertension in Adults: A Meta-Analysis of 7 646 267 Participants From 135 Studies. Journal of the American Heart Association. 2018;7(23).
64. Jornayvaz FR, Vollenweider P, Bochud M, Mooser V, Waeber G, Marques-Vidal P. Low birth weight leads to obesity, diabetes and increased leptin levels in adults: the CoLaus study. Cardiovascular Diabetology. 2016;15.
65. Zhao Y, Wang SF, Mu M, Sheng J. Birth weight and overweight/obesity in adults: a meta-analysis. European Journal of Pediatrics. 2012;171(12):1737-46.
66. Martin A, Connelly A, Bland RM, Reilly JJ. Health impact of catch-up growth in low-birth weight infants: systematic review, evidence appraisal, and meta-analysis. Maternal and Child Nutrition. 2017;13(1).
67. Magnus P, Gjessing HK, Skrondal A, Skjaerven R. Paternal contribution to birth weight. Journal of Epidemiology and Community Health. 2001;55(12):873-7.
68. McCowan LME, North RA, Kho EM, Black MA, Chan EHY, Dekker GA, et al. Paternal Contribution to Small for Gestational Age Babies: A Multicenter Prospective Study. Obesity. 2011;19(5):1035-9.

69. Hillman S, Peebles DM, Williams DJ. Paternal metabolic and cardiovascular risk factors for fetal growth restriction: a case-control study. Diabetes Care. 2013;36(6):1675-80.
70. Frayling TM, Hattersley AT. The role of genetic susceptibility in the association of low birth weight with type 2 diabetes. British Medical Bulletin. 2001;60:89-101.
71. Dickens LT, Letourneau LR, Sanyoura M, Greeley SAW, Philipson LH, Naylor RN. Management and pregnancy outcomes of women with GCK-MODY enrolled in the US Monogenic Diabetes Registry. Acta Diabetologica. 2019;56(4):405-11.
72. Hattersley AT, Beards F, Ballantyne E, Appleton M, Harvey R, Ellard S. Mutations in the glucokinase gene of the fetus result in reduced birth weight. Nature Genetics. 1998;19(3):268-70.
73. Horikoshi M, Beaumont RN, Day FR, Warrington NM, Kooijman MN, Fernandez-Tajes J, et al. Genome-wide associations for birth weight and correlations with adult disease. Nature. 2016;538(7624):248-+.
74. Beaumont RN, Horikoshi M, McCarthy MI, Freathy RM. How Can Genetic Studies Help Us to Understand Links Between Birth Weight and Type 2 Diabetes? Current Diabetes Reports. 2017;17(4).
75. Dupont C, Armant DR, Brenner CA. Epigenetics: Definition, Mechanisms and Clinical Perspective. Seminars in Reproductive Medicine. 2009;27(5):351-7.
76. Wu CT, Morris JR. Genes, genetics, and epigenetics: A correspondence. Science. 2001;293(5532):1103-5.
77. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30.
78. Tiffon C. The Impact of Nutrition and Environmental Epigenetics on Human Health and Disease. International Journal of Molecular Sciences. 2018;19(11).
79. Tost J. DNA Methylation Protocols Third Edition. Walker JM, editor. Evry, France: Humana Press; 2017.
80. Daxinger L, Whitelaw E. Understanding transgenerational epigenetic inheritance via the gametes in mammals. Nature Reviews Genetics. 2012;13(3):153-62.
81. Desai M, Jellyman JK, Ross MG. Epigenomics, gestational programming and risk of metabolic syndrome. Int J Obes (Lond). 2015;39(4):633-41.

82. Li E, Zhang Y. DNA Methylation in Mammals. Cold Spring Harbor Perspectives in Biology. 2014;6(5).
83. Dong XJ, Weng ZP. The correlation between histone modifications and gene expression. Epigenomics. 2013;5(2):113-6.
84. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nature Reviews Genetics. 2011;12(8):529-41.
85. Berdasco M, Esteller M. Clinical epigenetics: seizing opportunities for translation. Nature Reviews Genetics. 2019;20(2):109-27.
86. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nature Reviews Genetics. 2013;14(3):204-20.
87. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biology. 2016;17.
88. Breitling LP, Yang RX, Korn B, Burwinkel B, Brenner H. Tobacco-Smoking-Related Differential DNA Methylation: 27K Discovery and Replication. American Journal of Human Genetics. 2011;88(4):450-7.
89. Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, Beyan H, et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Research. 2010;20(4):434-9.
90. Rotroff DM, Joubert BR, Marvel SW, Haberg SE, Wu MC, Nilsen RM, et al. Maternal smoking impacts key biological pathways in newborns through epigenetic modification in Utero. Bmc Genomics. 2016;17.
91. Shen J, Wang S, Zhang YJ, Wu HC, Kibriya MG, Jasmine F, et al. Exploring genome-wide DNA methylation profiles altered in hepatocellular carcinoma using Infinium HumanMethylation 450 BeadChips. Epigenetics. 2013;8(1):34-43.
92. Heyn H, Carmona FJ, Gomez A, Ferreira HJ, Bell JT, Sayols S, et al. DNA methylation profiling in breast cancer discordant identical twins identifies DOK7 as novel epigenetic biomarker. Carcinogenesis. 2013;34(1):102-8.
93. Svendsen AJ, Gervin K, Lyle R, Christiansen L, Kyvik K, Junker P, et al. Differentially Methylated DNA Regions in Monozygotic Twin Pairs Discordant for Rheumatoid Arthritis: An Epigenome-Wide study. Frontiers in Immunology. 2016;7.
94. Visa N, Jordan-Pla A. ChIP and ChIP-Related Techniques: Expanding the Fields of Application and Improving ChIP Performance.

Chromatin Immunoprecipitation: Methods and Protocols. 2018;1689:1-7.
95. Onder O, Sidoli S, Carroll M, Garcia BA. Progress in epigenetic histone modification analysis by mass spectrometry for clinical investigations. Expert Review of Proteomics. 2015;12(5):499-517.
96. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009;10(1):57-63.
97. Youngson NA, Morris MJ. What obesity research tells us about epigenetic mechanisms. Philosophical Transactions of the Royal Society B-Biological Sciences. 2013;368(1609).
98. Tateishi K, Okada Y, Kallin EM, Zhang Y. Role of Jhdm2a in regulating metabolic gene expression and obesity resistance. Nature. 2009;458(7239):757-61.
99. van Dijk SJ, Molloy PL, Varinli H, Morrison JL, Muhlhausler BS, EpiScope. Epigenetics and human obesity. International Journal of Obesity. 2015;39(1):85-97.
100. Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aissi D, Wahl S, et al. DNA methylation and body-mass index: a genome-wide analysis. Lancet. 2014;383(9933):1990-8.
101. Aslibekyan S, Demerath EW, Mendelson M, Zhi D, Guan W, Liang L, et al. Epigenome-wide study identifies novel methylation loci associated with body mass index and waist circumference. Obesity. 2015;23(7):1493-501.
102. Demerath EW, Guan WH, Grove ML, Aslibekyan S, Mendelson M, Zhou YH, et al. Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. Human Molecular Genetics. 2015;24(15):4464-79.
103. Al Muftah WA, Al-Shafai M, Zaghlool SB, Visconti A, Tsai PC, Kumar P, et al. Epigenetic associations of type 2 diabetes and BMI in an Arab population. Clinical Epigenetics. 2016;8.
104. Mendelson MM, Marioni RE, Joehanes R, Liu CY, Hedman AK, Aslibekyan S, et al. Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach. Plos Medicine. 2017;14(1).
105. Sayols-Baixeras S, Subirana I, Fernandez-Sanles A, Senti M, Lluis-Ganella C, Marrugat J, et al. DNA methylation and obesity traits: An epigenome-wide association study. The REGICOR study. Epigenetics. 2017;12(10):909-16.

106. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017;541(7635):81-+.
107. Dhana K, Braun KVE, Nano J, Voortman T, Demerath EW, Guan W, et al. An Epigenome-Wide Association Study of Obesity-Related Traits. American Journal of Epidemiology. 2018;187(8):1662-9.
108. Wang XL, Pan Y, Zhu HD, Hao G, Huang YS, Barnes V, et al. An epigenome-wide study of obesity in African American youth and young adults: novel findings, replication in neutrophils, and relationship with gene expression. Clinical Epigenetics. 2018;10.
109. Pfeiffer S, Kruger J, Maierhofer A, Bottcher Y, Kloting N, El Hajj N, et al. Hypoxia-inducible factor 3A gene expression and methylation in adipose tissue is related to adipose tissue dysfunction. Scientific Reports. 2016;6.
110. Ling C, Ronn T. Epigenetics in Human Obesity and Type 2 Diabetes. Cell Metabolism. 2019;29(5):1028-44.
111. Eberle D, Hegarty B, Bossard P, Ferre P, Foufelle F. SREBP transcription factors: master regulators of lipid homeostasis. Biochimie. 2004;86(11):839-48.
112. Jelenik T, Kaul K, Sequaris G, Flogel U, Phielix E, Kotzka J, et al. Mechanisms of Insulin Resistance in Primary and Secondary Nonalcoholic Fatty Liver. Diabetes. 2017;66(8):2241-53.
113. Ruiz R, Jideonwo V, Ahn M, Surendran S, Tagliabracci VS, Hou YY, et al. Sterol Regulatory Element-binding Protein-1 (SREBP-1) Is Required to Regulate Glycogen Synthesis and Gluconeogenic Gene Expression in Mouse Liver. Journal of Biological Chemistry. 2014;289(9):5510-7.
114. Saxena R, Elbers CC, Guo YR, Peter I, Gaunt TR, Mega JL, et al. Large-Scale Gene-Centric Meta-Analysis across 39 Studies Identifies Type 2 Diabetes Loci. American Journal of Human Genetics. 2012;90(3):410-25.
115. Toperoff G, Aran D, Kark JD, Rosenberg M, Dubnikov T, Nissan B, et al. Genome-wide survey reveals predisposing diabetes type 2-related DNA methylation variations in human peripheral blood. Human Molecular Genetics. 2012;21(2):371-83.
116. Walaszczyk E, Luijten M, Spijkerman AMW, Bonder MJ, Lutgers HL, Snieder H, et al. DNA methylation markers associated with type 2 diabetes, fasting glucose and HbA(1c) levels: a systematic review and replication in a case-control sample of the Lifelines study. Diabetologia. 2018;61(2):354-68.

117. Kitade M, Ogura Y, Monno I, Koya D. Sirtuins and Type 2 Diabetes: Role in Inflammation, Oxidative Stress, and Mitochondrial Function. Frontiers in Endocrinology. 2019;10.
118. Zhou S, Tang XQ, Chen HZ. Sirtuins and Insulin Resistance. Frontiers in Endocrinology. 2018;9.
119. Yoshizaki T, Schenk S, Imamura T, Babendure JL, Sonoda N, Bae EJ, et al. SIRT1 inhibits inflammatory pathways in macrophages and modulates insulin sensitivity. American Journal of Physiology-Endocrinology and Metabolism. 2010;298(3):E419-E28.
120. Lorente-Cebrian S, Gonzalez-Muniesa P, Milagro FI, Martinez JA. MicroRNAs and other non-coding RNAs in adipose tissue and obesity: emerging roles as biomarkers and therapeutic targets. Clinical Science. 2019;133(1):23-40.
121. Zhao XY, Xiong XL, Liu TY, Mi L, Peng XL, Rui C, et al. Long noncoding RNA licensing of obesity-linked hepatic lipogenesis and NAFLD pathogenesis. Nature Communications. 2018;9.
122. Lo KA, Huang SQ, Walet ACE, Zhang ZC, Leow MKS, Liu MH, et al. Adipocyte Long-Noncoding RNA Transcriptome Analysis of Obese Mice Identified Lnc-Leptin, Which Regulates Leptin. Diabetes. 2018;67(6):1045-56.
123. Benn M, Nordestgaard BG. From genome-wide association studies to Mendelian randomization: novel opportunities for understanding cardiovascular disease causality, pathogenesis, prevention, and treatment. Cardiovascular Research. 2018;114(9):1192-208.
124. Hattersley AT, Tooke JE. The fetal insulin hypothesis: an alternative explanation of the association of low birthweight with diabetes and vascular disease. Lancet. 1999;353(9166):1789-92.
125. Knudsen TM, Rezwan FI, Jiang Y, Karmaus W, Svanes C, Holloway JW. Transgenerational and intergenerational epigenetic inheritance in allergic diseases. Journal of Allergy and Clinical Immunology. 2018;142(3):765-72.
126. Aiken CE, Ozanne SE. Transgenerational developmental programming. Human Reproduction Update. 2014;20(1):63-75.
127. Huypens P, Sass S, Wu M. Epigenetic germline inheritance of diet-induced obesity and insulin resistance. Nature Genetics. 2016.
128. Tang WWC, Dietmann S, Irie N, Leitch HG, Floros VI, Bradshaw CR, et al. A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development. Cell. 2015;161(6):1453-67.

129. Santiago M, Antunes C, Guedes M, Sousa N, Marques CJ. TET enzymes and DNA hydroxymethylation in neural development and function - How critical are they? Genomics. 2014;104(5):334-40.
130. Hackett JA, Sengupta R, Zylicz JJ, Murakami K, Lee C, Down TA, et al. Germline DNA Demethylation Dynamics and Imprint Erasure Through 5-Hydroxymethylcytosine. Science. 2013;339(6118):448-52.
131. Ge SQ, Lin SL, Zhao ZH, Sun QY. Epigenetic dynamics and interplay during spermatogenesis and embryogenesis: implications for male fertility and offspring health. Oncotarget. 2017;8(32):53804-18.
132. Barbosa TD, Ingerslev LR, Alm PS, Versteyhe S, Massart J, Rasmussen M, et al. High-fat diet reprograms the epigenome of rat spermatozoa and transgenerationally affects metabolism of the offspring. Molecular Metabolism. 2016;5(3):184-97.
133. Radford EJ, Ito M, Shi H, Corish JA, Yamazawa K, Isganaitis E, et al. In utero undernourishment perturbs the adult sperm methylome and intergenerational metabolism. Science. 2014;345(6198):785-+.
134. Sakai K, Ideta-Otsuka M, Saito H, Hiradate Y, Hara K, Igarashi K, et al. Effects of doxorubicin on sperm DNA methylation in mouse models of testicular toxicity. Biochemical and Biophysical Research Communications. 2018;498(3):674-9.
135. Dias BG, Ressier KJ. Parental olfactory experience influences behavior and neural structure in subsequent generations. Nature Neuroscience. 2014;17(1):89-96.
136. Watkins AJ, Dias I, Tsuro H, Allen D, Emes RD, Moreton J, et al. Paternal diet programs offspring health through sperm- and seminal plasma-specific pathways in mice. Proceedings of the National Academy of Sciences of the United States of America. 2018;115(40):10064-9.
137. Youngson NA, Lecomte V, Maloney CA, Leung P, Liu J, Hesson LB, et al. Obesity-induced sperm DNA methylation changes at satellite repeats are reprogrammed in rat offspring. Asian Journal of Andrology. 2016;18(6):930-6.
138. Ng SF, Lin RCY, Laybutt DR, Barres R, Owens JA, Morris MJ. Chronic high-fat diet in fathers programs beta-cell dysfunction in female rat offspring. Nature. 2010;467(7318):963-U103.
139. Huypens P, Sass S, Wu M, Dyckhoff D, Tschop M, Theis F, et al. Epigenetic germline inheritance of diet-induced obesity and insulin resistance. Nature Genetics. 2016;48(5):497-+.
140. Wei YC, Yang CR, Wei YP, Zhao ZA, Hou Y, Schatten H, et al. Paternally induced transgenerational inheritance of susceptibility to

diabetes in mammals. Proceedings of the National Academy of Sciences of the United States of America. 2014;111(5):1873-8.
141. Martinez D, Pentinat T, Ribo S, Daviaud C, Bloks VW, Cebria J, et al. In Utero Undernutrition in Male Mice Programs Liver Lipid Metabolism in the Second-Generation Offspring Involving Altered Lxra DNA Methylation. Cell Metabolism. 2014;19(6):941-51.
142. Horsthemke B. A critical view on transgenerational epigenetic inheritance in humans. Nature Communications. 2018;9.
143. Shea JM, Serra RW, Carone BR, Shulha HP, Kucukural A, Ziller MJ, et al. Genetic and Epigenetic Variation, but Not Diet, Shape the Sperm Methylome. Developmental Cell. 2015;35(6):750-8.
144. Sharma U, Conine CC, Shea JM, Boskovic A, Derr AG, Bing XY, et al. Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. Science. 2016;351(6271):391-6.
145. Chen Q, Yan MH, Cao ZH, Li X, Zhang YF, Shi JC, et al. Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. Science. 2016;351(6271):397-400.
146. Donkin I, Versteyhe S, Qian K, Ingerslev LR, Kantor FR, Mechta M, et al. Spermatozoa from lean and obese human carry distinct epigenetic signatures. Diabetologia. 2014;57:S105-S.
147. Donkin I, Versteyhe S, Ingerslev LR, Qian K, Mechta M, Nordkap L, et al. Obesity and Bariatric Surgery Drive Epigenetic Variation of Spermatozoa in Humans. Cell Metabolism. 2016;23(2):369-78.
148. Denham J, O'Brien BJ, Harvey JT, Charchar FJ. Genome-wide sperm DNA methylation changes after 3 months of exercise training in humans. Epigenomics. 2015;7(5):717-31.
149. Ingerslev LR, Donkin I, Fabre O, Versteyhe S, Mechta M, Pattamaprapanont P, et al. Endurance training remodels sperm-borne small RNA expression and methylation at neurological gene hotspots. Clinical Epigenetics. 2018;10.
150. Hillman SL, Finer S, Smart MC, Mathews C, Lowe R, Rakyan VK, et al. Novel DNA methylation profiles associated with key gene regulation and transcription pathways in blood and placenta of growth-restricted neonates. Epigenetics. 2015;10(1):50-61.
151. Engel SM, Joubert BR, Wu MC, Olshan AF, Haberg SE, Ueland PM, et al. Neonatal Genome-Wide Methylation Patterns in Relation to Birth Weight in the Norwegian Mother and Child Cohort. American Journal of Epidemiology. 2014;179(7):834-42.

152. Tobi EW, Goeman JJ, Monajemi R, Gu HC, Putter H, Zhang YJ, et al. DNA methylation signatures link prenatal famine exposure to growth and metabolism. Nature Communications. 2014;5:13.
153. Soubry A, Murphy SK, Wang F, Huang Z, Vidal AC, Fuemmeler BF, et al. Newborns of obese parents have altered DNA methylation patterns at imprinted genes. International Journal of Obesity. 2015;39(4):650-7.
154. Gosden R, Trasler J, Lucifero D, Faddy M. Rare congenital disorders, imprinted genes, and assisted reproductive technology. Lancet. 2003;361(9373):1975-7.
155. Ounap K. Silver-Russell Syndrome and Beckwith-Wiedemann Syndrome: Opposite Phenotypes with Heterogeneous Molecular Etiology. Molecular Syndromology. 2016;7(3):110-21.
156. Kalsner L, Chamberlain SJ. Prader-Willi, Angelman, and 15q11-q13 Duplication Syndromes. Pediatric Clinics of North America. 2015;62(3):587-+.
157. Boissonnas CC, El Abdalaoui H, Haelewyn V, Fauque P, Dupont JM, Gut I, et al. Specific epigenetic alterations of IGF2-H19 locus in spermatozoa from infertile men. European Journal of Human Genetics. 2010;18(1):73-80.
158. Camprubi C, Pladevall M, Grossmann M, Garrido N, Pons MC, Blanco J. Semen samples showing an increased rate of spermatozoa with imprinting errors have a negligible effect in the outcome of assisted reproduction techniques. Epigenetics. 2012;7(10):1115-24.
159. Manning M, Lissens W, Liebaers I, Van Steirteghem A, Weidner W. Imprinting analysis in spermatozoa prepared for intracytoplasmic sperm injection (ICSI). International Journal of Andrology. 2001;24(2):87-94.
160. Al-Khtib M, Blachere T, Guerin JF, Lefevre A. Methylation profile of the promoters of Nanog and Oct4 in ICSI human embryos. Human Reproduction. 2012;27(10):2948-54.
161. Navarro-Costa P, Nogueira P, Carvalho M, Leal F, Cordeiro I, Calhaz-Jorge C, et al. Incorrect DNA methylation of the DAZL promoter CpG island associates with defective human sperm(dagger). Human Reproduction. 2010;25(10):2647-54.
162. Rotondo JC, Bosi S, Bazzan E, Di Domenico M, De Mattei M, Selvatici R, et al. Methylenetetrahydrofolate reductase gene promoter hypermethylation in semen samples of infertile couples correlates with recurrent spontaneous abortion. Human Reproduction. 2012;27(12):3632-8.

163. Aston KI, Uren PJ, Jenkins TG, Horsager A, Cairns BR, Smith AD, et al. Aberrant sperm DNA methylation predicts male fertility status and embryo quality. Fertility and Sterility. 2015;104(6):1388-+.
164. Camprubi C, Salas-Huetos A, Aiese-Cigliano R, Godo A, Pons MC, Castellano G, et al. Spermatozoa from infertile patients exhibit differences of DNA methylation associated with spermatogenesis-related processes: an array-based analysis. Reproductive Biomedicine Online. 2016;33(6):709-19.
165. Jenkins TG, Aston KI, Meyer TD, Hotaling JM, Shamsi MB, Johnstone EB, et al. Decreased fecundity and sperm DNA methylation patterns. Fertility and Sterility. 2016;105(1):51-+.
166. Laqqan M, Solomayer EF, Hammadeh M. Association between alterations in DNA methylation level of spermatozoa at CpGs dinucleotide and male subfertility problems. Andrologia. 2018;50(1).
167. Laqqan M, Solomayer EF, Hammadeh M. Aberrations in sperm DNA methylation patterns are associated with abnormalities in semen parameters of subfertile males. Reproductive Biology. 2017;17(3):246-51.
168. Krausz C, Sandoval J, Sayols S, Chianese C, Giachini C, Heyn H, et al. Novel Insights into DNA Methylation Features in Spermatozoa: Stability and Peculiarities. Plos One. 2012;7(10).
169. Kobayashi H, Sato A, Otsu E, Hiura H, Tomatsu C, Utsunomiya T, et al. Aberrant DNA methylation of imprinted loci in sperm from oligospermic patients. Human Molecular Genetics. 2007;16(21):2542-51.
170. Roseboom T, de Rooij S, Painter R. The Dutch famine and its long-term consequences for adult health. Early Human Development. 2006;82(8):485-91.
171. Horikoshi M, Yaghootkar H, Mook-Kanamori DO, Sovio U, Taal HR, Hennig BJ, et al. New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. Nature Genetics. 2013;45(1):76-U115.
172. Machin D, Campbell MJ, Beng Tan S, Huey Tan S. Sample Size Tables for Clinical Studies, Third Edition: Wiley-Blackwell; 2009.
173. Norris T, Seaton SE, Manktelow BN, Baker PN, Kurinczuk JJ, Field D, et al. Updated birth weight centiles for England and Wales. Archives of Disease in Childhood-Fetal and Neonatal Edition. 2018;103(6):F577-F82.
174. which.co.uk Independent review site. The Labour Ward, Elizabeth Garrett Anderson Wing, UCH 2018 [updated 28 Jun 2018. Available

from: https://www.which.co.uk/birth-choice/maternity-units/ou-the-labour-ward-elizabeth-garrett-anderson-wing-uch-university-college-hospital-london#contacts.
175. Figueras F, Gratacos E. An integrated approach to fetal growth restriction. Best Practice & Research Clinical Obstetrics & Gynaecology. 2017;38:48-58.
176. The Perinatal Institute. Gestation Network Centile Calculator 2018 [Bulk centile calculator available upon request from the Perinatal Institute]. Available from: https://www.gestation.net/cc/about.htm.
177. Gardosi J, Giddings S, Buller S, Southam M, Williams M. Preventing stillbirths through improved antenatal recognition of pregnancies at risk due to fetal growth restriction. Public Health. 2014;128(8):698-702.
178. American Diabetes Association. Diagnosis and Classification of Diabetes Mellitus. Diabetes Care. 2009;32:S62-S7.
179. World Health Organization. Waist circumference and waist–hip ratioReport of a WHO expert consultation, Geneva, 8-11 December 2008. Report. Geneva: World Health Organisation; 2008 8-11 December 2008. Contract No.: ISBN : 978 92 4 150149 1.
180. Qiagen. QIAamp. DNA Mini and Blood Mini Handbook 1102728. Fifth edition ed: Qiagen HB-0329-004; May 2016.
181. Muniyappa R, Lee S, Chen H, Quon MJ. Current approaches for assessing insulin sensitivity and resistance in vivo: advantages, limitations, and appropriate usage. American Journal of Physiology-Endocrinology and Metabolism. 2008;294(1):E15-E26.
182. Wallace TM, Levy JC, Matthews DR. Use and abuse of HOMA modeling. Diabetes Care. 2004;27(6):1487-95.
183. Levy JC, Matthews DR, Hermans MP. Correct homeostasis model assessment (HOMA) evaluation uses the computer program. Diabetes Care. 1998;21(12):2191-2.
184. Oxford Centre for Diabetes Endocrinology and Metabolism. HOMA Calculator from the Diabetes Trials Unit: Oxford Centre for Diabetes, Endocrinology and Metabolism; 2004 [Available from: https://www.dtu.ox.ac.uk/homacalculator/.
185. Otten J, Ahren B, Olsson T. Surrogate measures of insulin sensitivity vs the hyperinsulinaemic-euglycaemic clamp: a meta-analysis. Diabetologia. 2014;57(9):1781-8.
186. National Institute for Health Research. Introduction to Good Clinical Practice eLearning (Secondary Care) 2017 [Available from: https://www.nihr.ac.uk/our-research-community/clinical-research-

staff/learning-and-development/national-directory/good-clinical-practice/our-courses/.
187. Medical Research Council. Use of human samples in medical research [e-Learning package]. 2004 [updated 2016. MRC guidelines and e-Learning package for the use of human tissue in medical research]. Available from: https://mrc.ukri.org/research/facilities-and-resources-for-researchers/regulatory-support-centre/human-tissue/.
188. Bartolomei MS, Ferguson-Smith AC. Mammalian Genomic Imprinting. Cold Spring Harbor Perspectives in Biology. 2011;3(7):17.
189. Qiagen. Qiagen Safety Data Sheet for Buffer AL-T/M Victoria 3148, Australia: Qiagen; 02.08.2018 [Available from: https://www.qiagen.com/au/products/human-identity-and-forensics/investigator-solutions/assay-setup/buffer-al/#resources.
190. Qiagen. Qiagen Resources: Buffer AE [Available from: https://www.qiagen.com/gb/resources/faq?id=c484a4ad-6f46-4cb5-96f3-875b72f41512&lang=en.
191. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. Bmj-British Medical Journal. 2009;339.
192. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53-9.
193. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. Nature. 2003;422(6934):835-47.
194. Ryan R, Hill S. How to GRADE the quality of the evidence. Cochrane Consumers and Communication Group La Trobe University, Melbourne2016 [3.0:[Available from: http://cccrg.cochrane.org/author-resources.
195. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic Signatures of Cigarette Smoking. Circulation-Cardiovascular Genetics. 2016;9(5):436-47.
196. Horvath S. DNA methylation age of human tissues and cell types. Genome Biology. 2013;14(10).
197. Tost Je. DNA Methylation Protocols Third Edition. Third ed. Walker JM, editor. Evry, France: Humana Press; 2017.
198. Huang WY, Hsu SD, Huang HY, Sun YM, Chou CH, Weng SL, et al. MethHC: a database of DNA methylation and gene expression in human cancer. Nucleic Acids Research. 2015;43(D1):D856-D61.

199. Illumina. Pub. No. 1070-2015-008-B. Infinium MethylationEPIC BeadChip Datasheet. Illumina; 2017.
200. Illumina. Infinium HD Assay Methylation Protocol Guide Document # 15019519 [PDF]: Illumina, Inc; 2015. Available from: http://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/infinium_assays/infinium_hd_methylation/infinium-hd-methylation-guide-15019519-01.pdf.
201. Illumina. Illumina Pub. No. 270-2012-001 Methylation BeadChips Achieve Breadth of Coverage Using 2 Infinium Chemistries Technical Note. Illumina; 2015.
202. Smith M, Baggerly K, Bengtsson H, Ritchie M, Hansen K. *illuminaio*: An open source IDAT parsing tool for Illumina microarrays. F1000Res. 2013.
203. Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. Bmc Genomics. 2013;14.
204. Andrews SV, Ladd-Acosta C, Feinberg AP, Hansen KD, Fallin MD. "Gap hunting" to characterize clustered probe signals in Illumina methylation array data. Epigenetics & Chromatin. 2016;9.
205. Carbon S, Dietze H, Lewis SE, Mungall CJ, Munoz-Torres MC, Basu S, et al. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Research. 2017;45(D1):D331-D8.
206. Dong D, YZheng, STeschendorff, A. ebGSEA: an improved Gene Set Enrichment Analysis method for Epigenome-Wide-Association Studies. Bioinformatics. 2019.
207. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics. 2016;32(2):286-8.
208. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.
209. Volkov P, Olsson AH, Gillberg L, Jorgensen SW, Brons C, Eriksson KF, et al. A Genome-Wide mQTL Analysis in Human Adipose Tissue Identifies Genetic Variants Associated with DNA Methylation, Gene Expression and Metabolic Traits. Plos One. 2016;11(6).
210. Clough E, Barrett T. The Gene Expression Omnibus Database. Statistical Genomics: Methods and Protocols. 2016;1418:93-110.

211. Gorrie-Stone TJ, Smart MC, Saffari A, Malki K, Hannon E, Burrage J, et al. Bigmelon: tools for analysing large DNA methylation datasets. Bioinformatics. 2019;35(6):981-6.
212. Voisin S, Almen MS, Zheleznyakova GY, Lundberg L, Zarei S, Castillo S, et al. Many obesity-associated SNPs strongly associate with DNA methylation changes at proximal promoters and enhancers. Genome Medicine. 2015;7.
213. Ligthart S, Marzi C, Aslibekyan S, Mendelson MM, Conneely KN, Tanaka T, et al. DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. Genome Biology. 2016;17.
214. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. Bmc Bioinformatics. 2012;13.
215. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3-new capabilities and interfaces. Nucleic Acids Research. 2012;40(15).
216. GmbH RD. FastStart High Fidelity PCR System Mannheim, Germany: Roche Applied Science 68298; 2011 [0312. 03707555001]. Available from: https://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Roche/Bulletin/1/fhifirobul.pdf.
217. Qiagen. HotStarTaq ® PCR Handbook 1051481 Hilden, Germany: Qiagen; 2008 [Available from: https://www.hain-lifescience.de/uploadfiles/file/downloadcenter/polymerase/manual_polymerase.pdf.
218. Illumina. Access Array System for Illumina Sequencing Systems USER GUIDE PN 100-3770 M1 United States: Illumina; 2019 [Available from: https://www.fluidigm.com/binaries/content/documents/fluidigm/resources/aa-illumina-100-3770/aa-illumina-100-3770/fluidigm%3Afile.
219. Adamowicz M, Maratou K, Aitman T. Multiplexed DNA Methylation Analysis of Target Regions Using Microfluidics (Fluidigm). Chapter 18 in DNA Methylation Protocols , Methods in Molecular Biology. Tost J, editor. Evry, France: Springer Science+Business Media, LLC 2018; 2017.
220. Illumina. Illumina MiSeq Overview San Diego, California: Illumina; 2019 [Available from: https://emea.illumina.com/systems/sequencing-platforms/miseq.html.
221. Illumina. An introduction to Next-Generation Sequencing Technology San Diego, California: Illumina; 2017 [Available from:

https://www.illumina.com/documents/products/illumina_sequencing_introduction.pdf.
222. Sharma D, Farahbakhsh N, Shastri S, Sharma P. Intrauterine growth restriction – part 1. The Journal of Maternal-Fetal & Neonatal Medicine. 2016.
223. Lecomte V, Maloney CA, Wang KW, Morris MJ. Effects of paternal obesity on growth and adiposity of male rat offspring. American Journal of Physiology-Endocrinology and Metabolism. 2017;312(2):E117-E25.
224. Davey Smith G, Sterne J, Tynelius P, Rasmussen F. Birth characteristics of offspring and parental diabetes: evidence for the fetal insulin hypothesis. Journal of Epidemiology and Community Health. 2004.
225. Hypponen E, Smith GD, Power C. Parental diabetes and birth weight of offspring: intergenerational cohort study. British Medical Journal. 2003;326(7379):19-20.
226. Wannamethee SG, Lawlor DA, Whincup PH, Walker M, Ebrahim S, Davey-Smith G. Birthweight of offspring and paternal insulin resistance and paternal diabetes in late adulthood: cross sectional survey. Diabetologia. 2004;47(1):12-8.
227. Oldereid NB, Wennerholm UB, Pinborg A, Loft A, Laivuori H, Petzold M, et al. The effect of paternal factors on perinatal and paediatric outcomes: a systematic review and meta-analysis. Human Reproduction Update. 2018;24(3):320-89.
228. Gardosi J, Francis A, Turner S, Williams M. Customized growth charts: rationale, validation and clinical benefits. American Journal of Obstetrics and Gynecology. 2018;218(2):S609-S18.
229. Watterberg KL, Aucott S, Benitz WE, Cummings JJ, Eichenwald EC, Goldsmith J, et al. The Apgar Score. Pediatrics. 2015;136(4):819-22.
230. UCLH Clinical Biochemistry. UCLH Clinical Biochemistry Test Information University College London Hospital2017 [Biochemistry test information]. Available from: https://www.uclh.nhs.uk/OurServices/ServiceA-Z/PATH/PATHBIOMED/CBIO/Pages/InformationforGPs.aspx.
231. Gayoso-Diz P, Otero-Gonzalez A, Rodriguez-Alvarez MX, Gude F, Garcia F, De Francisco A, et al. Insulin resistance (HOMA-IR) cut-off values and the metabolic syndrome in a general adult population: effect of gender and age: EPIRCE cross-sectional study. Bmc Endocrine Disorders. 2013;13.
232. Geloneze B, Vasques AC, Stabe CF, Rosado LE, Queiroz EC, Tambascia MA, et al. Homeostatic Model Assessment Indexes (HOMA1-

IR and HOMA2-IR) To Identify Insulin Resistance and Metabolic Syndrome in an Admixtured Multiethnic Population-BRAMS-Brazilian Metabolic Syndrome Study. Diabetes. 2009;58:A260-A.
233. National Heart L, and Blood Institute (US). The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. Bethesda: U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES; 2004.
234. Sliwinska-Mosson M, Milnerowicz H. The impact of smoking on the development of diabetes and its complications. Diabetes & Vascular Disease Research. 2017;14(4):265-76.
235. Villar J, Ismail LC, Victora CG, Ohuma EO, Bertino E, Altman DG, et al. International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project. Lancet. 2014;384(9946):857-68.
236. Anderson NH, Sadler LC, McKinlay CJD, McCowan LME. INTERGROWTH-21st vs customized birthweight standards for identification of perinatal mortality and morbidity. American Journal of Obstetrics and Gynecology. 2016;214(4).
237. Chen Y-P, Xiao X-M, Li J, Reichetzeder C, Wang Z-N, Hocher B. Paternal Body Mass Index (BMI) Is Associated with Offspring Intrauterine Growth in a Gender Dependent Manner. Plos One. 2012;7(5).
238. Radford EJ, Ito M, Shi H, Corish JA, Yamazawa K, Isganaitis E, et al. In utero undernourishment perturbs the adult sperm methylome and intergenerational metabolism. Science. 2014;345(6198):785-+.
239. Fields E, Chard J, James D, Treasure T, Guideline Dev G. GUIDELINES Fertility (update): summary of NICE guidance. Bmj-British Medical Journal. 2013;346.
240. Amer Soc Reprod M. Definitions of infertility and recurrent pregnancy loss: a committee opinion. Fertility and Sterility. 2013;99(1):63-.
241. Guzick DS, Overstreet JW, Factor-Litvak P, Brazil CK, Nakajima ST, Coutifaris C, et al. Sperm morphology, motility, and concentration in fertile and infertile men. New England Journal of Medicine. 2001;345(19):1388-93.
242. Ferlin A, Arredi B, Foresta C. Genetic causes of male infertility. Reproductive Toxicology. 2006;22(2):133-41.
243. Mima M, Greenwald D, Ohlander S. Environmental Toxins and Male Fertility. Current Urology Reports. 2018;19(7).

244. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. British Medical Journal. 2008;336(7650):924-6.
245. World Health Organization. WHO laboratory manual for the examination and processing of human semen- Fifth Edition. WHO, editor. Geneva, Switzerland: WHO; 2010.
246. Chowdhury B, Cho IH, Irudayaraj J. Technical advances in global DNA methylation analysis in human cancers. Journal of Biological Engineering. 2017;11.
247. Benchaib M, Ajina M, Lornage J, Niveleau A, Durand P, Guerin JF. Quantitation by image analysis of global DNA methylation in human spermatozoa and its prognostic value in in vitro fertilization: a preliminary study. Fertility and Sterility. 2003;80(4):947-53.
248. Singh RK, Diaz PE, Binette F, Nasonkin IO. Immunohistochemical Detection of 5-Methylcytosine and 5-Hydroxymethylcytosine in Developing and Postmitotic Mouse Retina. Jove-Journal of Visualized Experiments. 2018(138).
249. Barzideh J, Scott RJ, Aitken RJ. Analysis of the global methylation status of human spermatozoa and its association with the tendency of these cells to enter apoptosis. Andrologia. 2013;45(6):424-9.
250. Benchaib M, Braun V, Ressnikof D, Lornage J, Durand P, Niveleau A, et al. Influence of global sperm DNA methylation on IVF results. Human Reproduction. 2005;20(3):768-73.
251. Aoki VW, Emery BR, Carrell DT. Global sperm deoxyribonucleic acid methylation is unaffected in protamine-deficient infertile males. Fertility and Sterility. 2006;86(5):1541-3.
252. Tavalaee M, Razavi S, Nasr-Esfahani MH. Influence of sperm chromatin anomalies on assisted reproductive technology outcome. Fertility and Sterility. 2009;91(4):1119-26.
253. Tunc O, Tremellen K. Oxidative DNA damage impairs global sperm DNA methylation in infertile men. Journal of Assisted Reproduction and Genetics. 2009;26(9-10):537-44.
254. Jenkins TG, Aston KI, Cairns BR, Carrell DT. Paternal aging and associated intraindividual alterations of global sperm 5-methylcytosine and 5-hydroxymethylcytosine levels. Fertility and Sterility. 2013;100(4):945-+.
255. Leter G, Consales C, Eleuteri P, Uccelli R, Specht IO, Toft G, et al. Exposure to Perfluoroalkyl Substances and Sperm DNA Global

Methylation in Arctic and European Populations. Environmental and Molecular Mutagenesis. 2014;55(7):591-600.
256. Consales C, Leter G, Bonde JPE, Toft G, Eleuteri P, Moccia T, et al. Indices of methylation in sperm DNA from fertile men differ between distinct geographical regions. Human Reproduction. 2014;29(9):2065-72.
257. Bahreinian M, Tavalaee M, Abbasi H, Kiani-Esfahani A, Shiravi AH, Nasr-Esfahani MH. DNA hypomethylation predisposes sperm to DNA damage in individuals with varicocele. Systems Biology in Reproductive Medicine. 2015;61(4):179-86.
258. Yu BL, Zhou H, Liu M, Zheng T, Jiang L, Zhao M, et al. Epigenetic Alterations in Density Selected Human Spermatozoa for Assisted Reproduction. Plos One. 2015;10(12).
259. Kim SK, Jee BC, Kim SH. Histone methylation and acetylation in ejaculated human sperm: effects of swim-up and smoking. Fertility and Sterility. 2015;103(6):1425-31.
260. Tavalaee M, Bahreinian M, Barekat F, Abbasi H, Nasr-Esfahani MH. Effect of varicocelectomy on sperm functional characteristics and DNA methylation. Andrologia. 2015;47(8):904-9.
261. Cassuto NG, Montjean D, Siffroi JP, Bouret D, Marzouk F, Copin H, et al. Different Levels of DNA Methylation Detected in Human Sperms after Morphological Selection Using High Magnification Microscopy. Biomed Research International. 2016.
262. Consales C, Toft G, Leter G, Bonde JPE, Uccelli R, Pacchierotti F, et al. Exposure to persistent organic pollutants and sperm DNA methylation changes in Arctic and European populations. Environmental and Molecular Mutagenesis. 2016;57(3):200-9.
263. Uppangala S, Mathai SE, Salian SR, Kumar D, Singh VJ, D'Souza F, et al. Sperm Chromatin Immaturity Observed in Short Abstinence Ejaculates Affects DNA Integrity and Longevity In Vitro. Plos One. 2016;11(4).
264. Olszewska M, Barciszewska MZ, Fraczek M, Huleyuk N, Chernykh VB, Zastavna D, et al. Global methylation status of sperm DNA in carriers of chromosome structural aberrations. Asian Journal of Andrology. 2017;19(1):117-24.
265. Ren XX, Chen XL, Wang ZL, Wang D. Is transcription in sperm stationary or dynamic? Journal of Reproduction and Development. 2017;63(5):439-43.
266. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics. 2012;13(7):484-92.

267. Li YF, Sasaki H. Genomic imprinting in mammals: its life cycle, molecular mechanisms and reprogramming. Cell Research. 2011;21(3):466-73.
268. Jong MTC, Gray TA, Ji YG, Glenn CC, Saitoh S, Driscoll DJ, et al. A novel imprinted gene, encoding a RING zinc-finger protein, and overlapping antisense transcript in the Prader-Willi syndrome critical region. Human Molecular Genetics. 1999;8(5):783-93.
269. El-Maarri O, Buiting K, Peery EG, Kroisel PM, Balaban B, Wagner K, et al. Maternal methylation imprints on human chromosome 15 are established during or after fertilization. Nature Genetics. 2001;27(3):341-4.
270. Hamatani T, Sasaki H, Ishihara K, Hida N, Maruyama T, Yoshimura Y, et al. Epigenetic mark sequence of the H19 gene in human sperm. Biochimica Et Biophysica Acta-Gene Structure and Expression. 2001;1518(1-2):137-44.
271. Manning M, Lissens W, Weidner W, Liebaers I. DNA methylation analysis in immature testicular sperm cells at different developmental stages. Urologia Internationalis. 2001;67(2):151-5.
272. Marques CJ, Carvalho F, Sousa M, Barros A. Genomic imprinting in disruptive spermatogenesis. Lancet. 2004;363(9422):1700-2.
273. Grunau C, Sanchez C, Ehrlich M, van der Bruggen P, Hindermann W, Rodriguez C, et al. Frequent DNA hypomethylation of human juxtacentromeric BAGE loci in cancer. Genes Chromosomes & Cancer. 2005;43(1):11-24.
274. Stoger R. In Vivo Methylation Patterns of the Leptin Promoter in Human and Mouse. Epigenetics. 2006;1(4):155-62.
275. Li ZX, Ma X, Wang ZH. A differentially methylated region of the DAZ1 gene in spermatic and somatic cells. Asian Journal of Andrology. 2006;8(1):61-7.
276. Geuns E, De Temmerman N, Hilven P, Van Steirteghem A, Liebaers I, De Rycke M. Methylation analysis of the intergenic differentially methylated region of DLK1-GTL2 in human. European Journal of Human Genetics. 2007;15(3):352-61.
277. Marques CJ, Costa P, Vaz B, Carvalho F, Fernandes S, Barros A, et al. Abnormal methylation of imprinted genes in human sperm is associated with oligozoospermia. Molecular Human Reproduction. 2008;14(2):67-73.
278. Poplinski A, Tuttelmann F, Kanber D, Horsthemke B, Gromoll J. Idiopathic male infertility is strongly associated with aberrant

methylation of MEST and IGF2/H19 ICR1. International Journal of Andrology. 2010;33(4):642-9.
279. Kobayashi H, Hiura H, John RM, Sato A, Otsu E, Kobayashi N, et al. DNA methylation errors at imprinted loci after assisted conception originate in the parental sperm. European Journal of Human Genetics. 2009;17(12):1582-91.
280. Marques CJ, Francisco T, Sousa S, Carvalho F, Barros A, Sousa M. Methylation defects of imprinted genes in human testicular spermatozoa. Fertility and Sterility. 2010;94(2):585-94.
281. Wu W, Shen OX, Qin YF, Niu XB, Lu CC, Xia YK, et al. Idiopathic Male Infertility Is Strongly Associated with Aberrant Promoter Methylation of Methylenetetrahydrofolate Reductase (MTHFR). Plos One. 2010;5(11).
282. Hammoud SS, Purwar J, Pflueger C, Cairns BR, Carrell DT. Alterations in sperm DNA methylation patterns at imprinted loci in two classes of infertility. Fertility and Sterility. 2010;94(5):1728-33.
283. Nanassy L, Carrell DT. Abnormal methylation of the promoter of CREM is broadly associated with male factor infertility and poor sperm quality but is improved in sperm selected by density gradient centrifugation. Fertility and Sterility. 2011;95(7):2310-4.
284. Minor A, Chow V, Ma S. Aberrant DNA methylation at imprinted genes in testicular sperm retrieved from men with obstructive azoospermia and undergoing vasectomy reversal. Reproduction. 2011;141(6):749-57.
285. Chorbov V, Todorov A, Lynskey M, Cicero T. Elevated levels of DNA methylation at the OPRM1 promoter in blood and sperm from male opioid addicts. J Opioid Manag. 2011;Jul-Aug;7(4):258-64.
286. Nanassy L, Carrell DT. Analysis of the methylation pattern of six gene promoters in sperm of men with abnormal protamination. Asian Journal of Andrology. 2011;13(2):342-6.
287. Ibala-Romdhane S, Al-Khtib M, Khoueiry R, Blachere T, Guerin JF, Lefevre A. Analysis of H19 methylation in control and abnormal human embryos, sperm and oocytes. European Journal of Human Genetics. 2011;19(11):1138-43.
288. Nettersheim D, Biermann K, Gillis AJM, Steger K, Looijenga LHJ, Schorle H. NANOG promoter methylation and expression correlation during normal and malignant human germ cell development. Epigenetics. 2011;6(1):114-22.
289. Sato A, Hiura H, Okae H, Miyauchi N, Abe Y, Utsunomiya T, et al. Assessing loss of imprint methylation in sperm from subfertile men

using novel methylation polymerase chain reaction Luminex analysis. Fertility and Sterility. 2011;95(1):129-34.
290. El Hajj N, Zechner U, Schneider E, Tresch A, Gromoll J, Hahn T, et al. Methylation Status of Imprinted Genes and Repetitive Elements in Sperm DNA from Infertile Males. Sexual Development. 2011;5(2):60-9.
291. Ankolkar M, Patil A, Warke H, Salvi V, Mokashi NK, Pathak S, et al. Methylation analysis of idiopathic recurrent spontaneous miscarriage cases reveals aberrant imprinting at H19 ICR in normozoospermic individuals. Fertility and Sterility. 2012;98(5):1186-92.
292. Kaminsky Z, Tochigi M, Jia P, Pal M, Mill J, Kwan A, et al. A multi-tissue analysis identifies HLA complex group 9 gene methylation differences in bipolar disorder. Molecular Psychiatry. 2012;17(7):728-40.
293. Klaver R, Bleiziffer A, Redmann K, Mallidis C, Kliesch S, Gromoll J. Routine cryopreservation of spermatozoa is safe - Evidence from the DNA methylation pattern of nine spermatozoa genes. Journal of Assisted Reproduction and Genetics. 2012;29(9):943-50.
294. Berthaut I, Montjean D, Dessolle L, Morcel K, Deluen F, Poirot C, et al. Effect of temozolomide on male gametes: an epigenetic risk to the offspring? Journal of Assisted Reproduction and Genetics. 2013;30(6):827-33.
295. Rotondo JC, Selvatici R, Di Domenico M, Marci R, Vesce F, Tognon M, et al. Methylation loss at H19 imprinted gene correlates with methylenetetrahydrofolate reductase gene promoter hypermethylation in semen samples from infertile males. Epigenetics. 2013;8(9):990-7.
296. Li B, Li JB, Xiao XF, Ma YF, Wang J, Liang XX, et al. Altered DNA Methylation Patterns of the H19 Differentially Methylated Region and the DAZL Gene Promoter Are Associated with Defective Human Sperm. Plos One. 2013;8(8).
297. Klaver R, Tuttelmann F, Bleiziffer A, Haaf T, Kliesch S, Gromoll J. DNA methylation in spermatozoa as a prospective marker in andrology. Andrology. 2013;1(5):731-40.
298. Montjean D, Ravel C, Benkhalifa M, Cohen-Bacrie P, Berthaut I, Bashamboo A, et al. Methylation changes in mature sperm deoxyribonucleic acid from oligozoospermic men: assessment of genetic variants and assisted reproductive technology outcome. Fertility and Sterility. 2013;100(5):1241-+.
299. Camprubi C, Pladevall M, Grossmann M, Garrido N, Pons MC, Blanco J. Lack of association of MTHFR rs1801133 polymorphism and

CTCFL mutations with sperm methylation errors in infertile patients. Journal of Assisted Reproduction and Genetics. 2013;30(9):1125-31.
300. Miao M, Zhou X, Li Y, Zhang O, Zhou Z, Li T, et al. LINE-1 hypomethylation in spermatozoa is associated with Bisphenol A exposure. Andrology. 2014;2(1):138-44.
301. Botezatu A, Socolov R, Socolov D, Iancu IV, Anton G. Methylation pattern of methylene tetrahydrofolate reductase and small nuclear ribonucleoprotein polypeptide N promoters in oligoasthenospermia: a case-control study. Reproductive Biomedicine Online. 2014;28(2):225-31.
302. Richardson ME, Bleiziffer A, Tuttelmann F, Gromoll J, Wilkinson MF. Epigenetic regulation of the RHOX homeobox gene cluster and its association with human male infertility. Human Molecular Genetics. 2014;23(1):12-23.
303. Kuhtz J, Schneider E, El Hajj N, Zimmermann L, Fust O, Linek B, et al. Epigenetic heterogeneity of developmentally important genes in human sperm: Implications for assisted reproduction outcome. Epigenetics. 2014;9(12):1648-58.
304. Guardiola M, Oliva I, Guillaumet A, Martin-Trujillo A, Rosales R, Vallve JC, et al. Tissue-specific DNA methylation profiles regulate liver-specific expression of the APOA1/C3/A4/A5 cluster and can be manipulated with demethylating agents on intestinal cells. Atherosclerosis. 2014;237(2):528-35.
305. Zhou JH, Zhou QZ, Lyu XM, Zhu T, Chen ZJ, Chen MK, et al. The Expression of Cysteine-Rich Secretory Protein 2 (CRISP2) and Its Specific Regulator miR-27b in the Spermatozoa of Patients with Asthenozoospermia. Biology of Reproduction. 2015;92(1).
306. Laurentino S, Beygo J, Nordhoff V, Kliesch S, Wistuba J, Borgmann J, et al. Epigenetic germline mosaicism in infertile men. Human Molecular Genetics. 2015;24(5):1295-304.
307. Tian MP, Bao HQ, Martin FL, Zhang J, Liu LP, Huang QY, et al. Association of DNA Methylation and Mitochondrial DNA Copy Number with Human Semen Quality (vol 91, 101, 2014). Biology of Reproduction. 2015;92(3).
308. Montjean D, Zini A, Ravel C, Belloc S, Dalleac A, Copin H, et al. Sperm global DNA methylation level: association with semen parameters and genome integrity. Andrology. 2015;3(2):235-40.
309. Zhang T, Zhang L, Yuan QG, Wang XL, Zhang Y, Wang JH, et al. The Noninvasive Detection of RAR beta 2 Promoter Methylation for the

Diagnosis of Prostate Cancer. Cell Biochemistry and Biophysics. 2015;71(2):925-30.
310. Soubry A, Guo LS, Huang ZQ, Hoyo C, Romanus S, Price T, et al. Obesity-related DNA methylation at imprinted genes in human sperm: Results from the TIEGER study. Clinical Epigenetics. 2016;8.
311. Wu CL, Ding XF, Tan HP, Li HG, Xiong CL. Alterations of testis-specific promoter methylation in cell-free seminal deoxyribonucleic acid of idiopathic nonobstructive azoospermic men with different testicular phenotypes. Fertility and Sterility. 2016;106(6):1331-7.
312. Zhang CW, Xue P, Gao LZ, Chen X, Lin KB, Yang XY, et al. Highly conserved epigenetic regulation of BOULE and DAZL is associated with human fertility. Faseb Journal. 2016;30(10):3424-40.
313. Xu J, Zhang A, Zhang Z, Wang P, Qian Y, He L, et al. DNA methylation levels of imprinted and nonimprinted genes DMRs associated with defective human spermatozoa. Andrologia. 2016;48(9):939-47.
314. Louie K, Minor A, Ng R, Poon K, Chow V, Ma S. Evaluation of DNA methylation at imprinted DMRs in the spermatozoa of oligozoospermic men in association with MTHFR C677T genotype. Andrology. 2016;4(5):825-31.
315. Atsem S, Reichenbach J, Potabattula R, Dittrich M, Nava C, Depienne C, et al. Paternal age effects on sperm FOXK1 and KCNA7 methylation and transmission into the next generation. Human Molecular Genetics. 2016;25(22):4996-5005.
316. Buckley L, Lacey M, Ehrlich M. Epigenetics of the myotonic dystrophy-associated DMPK gene neighborhood. Epigenomics. 2016;8(1):13-31.
317. Soubry A, Hoyo C, Butt CM, Fieuws S, Price TM, Murphy SK, et al. Human exposure to flame-retardants is associated with aberrant DNA methylation at imprinted genes in sperm. Environmental Epigenetics. 2017;3(1).
318. Dong H, Wang YX, Zou ZK, Chen LM, Shen CY, Xu SQ, et al. Abnormal Methylation of Imprinted Genes and Cigarette Smoking: Assessment of Their Association With the Risk of Male Infertility. Reproductive Sciences. 2017;24(1):114-23.
319. Kobayashi N, Miyauchi N, Tatsuta N, Kitamura A, Okae H, Hiura H, et al. Factors associated with aberrant imprint methylation and oligozoospermia. Scientific Reports. 2017;7.

320. Marques PI, Fernandes S, Carvalho F, Barros A, Sousa M, Marques CJ. DNA methylation imprinting errors in spermatogenic cells from maturation arrest azoospermic patients. Andrology. 2017;5(3):451-9.
321. Nasri F, Gharesi-Fard B, Jahromi BN, Farazi-fard MA, Banaei M, Davari M, et al. Sperm DNA methylation of H19 imprinted gene and male infertility. Andrologia. 2017;49(10).
322. Lu ZX, Ma YF, Gao LY, Li YJ, Li Q, Qiang M. Urine mercury levels correlate with DNA methylation of imprinting gene H19 in the sperm of reproductive-aged men. Plos One. 2018;13(4).
323. Gunes S, Agarwal A, Henkel R, Mahmutoglu AM, Sharma R, Esteves SC, et al. Association between promoter methylation of MLH1 and MSH2 and reactive oxygen species in oligozoospermic menA pilot study. Andrologia. 2018;50(3).
324. Tian MP, Liu LP, Zhang J, Huang QY, Shen HQ. Positive association of low-level environmental phthalate exposure with sperm motility was mediated by DNA methylation: A pilot study. Chemosphere. 2019;220:459-67.
325. Darbandi M, Darbandi S, Agarwal A, Baskaran S, Dutta S, Sengupta P, et al. Reactive oxygen species-induced alterations in H19-Igf2 methylation patterns, seminal plasma metabolites, and semen quality. Journal of Assisted Reproduction and Genetics. 2019;36(2):241-53.
326. Molloy AM, Pangilinan F, Brody LC. Genetic Risk Factors for Folate-Responsive Neural Tube Defects. Annual Review of Nutrition, Vol 37. 2017;37:269-91.
327. de Kluiver H, Buizer-Voskamp JE, Dolan CV, Boomsma DI. Paternal Age and Psychiatric Disorders: A Review. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics. 2017;174(3):202-13.
328. Flanagan JM, Popendikyte V, Pozdniakovaite N, Sobolev M, Assadzadeh A, Schumacher A, et al. Intra- and interindividual epigenetic variation in human germ cells. American Journal of Human Genetics. 2006;79(1):67-84.
329. Houshdaran S, Cortessis VK, Siegmund K, Yang A, Laird PW, Sokol RZ. Widespread Epigenetic Abnormalities Suggest a Broad DNA Methylation Erasure Defect in Abnormal Human Sperm. Plos One. 2007;2(12).
330. Rakyan VK, Down TA, Thorne NP, Flicek P, Kulesha E, Graf S, et al. An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). Genome Research. 2008;18(9):1518-29.

331. Fan SC, Zhang XG. CpG island methylation pattern in different human tissues and its correlation with gene expression. Biochemical and Biophysical Research Communications. 2009;383(4):421-5.
332. Hammoud SS, Nix DA, Zhang HY, Purwar J, Carrell DT, Cairns BR. Distinctive chromatin in human sperm packages genes for embryo development. Nature. 2009;460(7254):473-U47.
333. Zeschnigk M, Martin M, Betzl G, Kalbe A, Sirsch C, Buiting K, et al. Massive parallel bisulfite sequencing of CG-rich DNA fragments reveals that methylation of many X-chromosomal CpG islands in female blood DNA is incomplete. Human Molecular Genetics. 2009;18(8):1439-48.
334. Molaro A, Hodges E, Fang F, Song Q, McCombie WR, Hannon GJ, et al. Sperm Methylation Profiles Reveal Features of Epigenetic Inheritance and Evolution in Primates. Cell. 2011;146(6):1028-40.
335. Pacheco SE, Houseman EA, Christensen BC, Marsit CJ, Kelsey KT, Sigman M, et al. Integrative DNA Methylation and Gene Expression Analyses Identify DNA Packaging and Epigenetic Regulatory Genes Associated with Low Motility Sperm. Plos One. 2011;6(6).
336. Aston KI, Punj V, Liu LH, Carrell DT. Genome-wide sperm deoxyribonucleic acid methylation is altered in some men with abnormal chromatin packaging or poor in vitro fertilization embryogenesis. Fertility and Sterility. 2012;97(2):285-U327.
337. Li J, Harris RA, Cheung SW, Coarfa C, Jeong M, Goodell MA, et al. Genomic Hypomethylation in the Human Germline Associates with Selective Structural Mutability in the Human Genome. Plos Genetics. 2012;8(5).
338. Schutte B, El Hajj N, Kuhtz J, Nanda I, Gromoll J, Hahn T, et al. Broad DNA methylation changes of spermatogenesis, inflammation and immune response-related genes in a subgroup of sperm samples for assisted reproduction. Andrology. 2013;1(6):822-9.
339. Kim R, Kulkarni P, Hannenhalli S. Derepression of Cancer/Testis Antigens in cancer is associated with distinct patterns of DNA Hypomethylation. Bmc Cancer. 2013;13.
340. Jenkins TG, Aston KI, Pflueger C, Cairns BR, Carrell DT. Age-Associated Sperm DNA Methylation Alterations: Possible Implications in Offspring Disease Susceptibility. Plos Genetics. 2014;10(7).
341. Zeng J, Nagrajan HK, Yi SV. Fundamental diversity of human CpG islands at multiple biological levels. Epigenetics. 2014;9(4):483-91.
342. Feinberg JI, Bakulski KM, Jaffe AE, Tryggvadottir R, Brown SC, Goldman LR, et al. Paternal sperm DNA methylation associated with

early signs of autism risk in an autism-enriched cohort. International Journal of Epidemiology. 2015;44(4):1199-210.
343. Urdinguio RG, Bayon GF, Dmitrijeva M, Torano EG, Bravo C, Fraga MF, et al. Aberrant DNA methylation patterns of spermatozoa in men with unexplained infertility. Human Reproduction. 2015;30(5):1014-28.
344. Jenkins TG, Aston KI, Trost C, Farley J, Hotaling JM, Carrell DT. Intra-sample heterogeneity of sperm DNA methylation. Molecular Human Reproduction. 2015;21(4):313-9.
345. Aarabi M, Gabriel MCS, Chan D, Behan NA, Caron M, Pastinen T, et al. High-dose folic acid supplementation alters the human sperm methylome and is influenced by the MTHFR C677T polymorphism. Human Molecular Genetics. 2015;24(22):6301-13.
346. Donkin I, Versteyhe S, Ingerslev LR, Qian K, Mechta M, Nordkap L, et al. Obesity and Bariatric Surgery Drive Epigenetic Variation of Spermatozoa in Humans. Cell metabolism. 2016;23(2):369-78.
347. Du Y, Li MY, Chen J, Duan YG, Wang XB, Qiu Y, et al. Promoter targeted bisulfite sequencing reveals DNA methylation profiles associated with low sperm motility in asthenozoospermia. Human Reproduction. 2016;31(1):24-33.
348. Jenkins TG, Aston KI, Hotaling JM, Shamsi MB, Simon L, Carrell DT. Teratozoospermia and asthenozoospermia are associated with specific epigenetic signatures. Andrology. 2016;4(5):843-9.
349. Dere E, Huse S, Hwang K, Sigman M, Boekelheide K. Intra- and inter-individual differences in human sperm DNA methylation. Andrology. 2016;4(5):832-42.
350. Wu HT, Estill MS, Shershebnev A, Suvorov A, Krawetz SA, Whitcomb BW, et al. Preconception urinary phthalate concentrations and sperm DNA methylation profiles among men undergoing IVF treatment: a cross-sectional study. Human Reproduction. 2017;32(11):2159-69.
351. Shnorhavorian M, Schwartz SM, Stansfeld B, Sadler-Riggleman I, Beck D, Skinner MK. Differential DNA Methylation Regions in Adult Human Sperm following Adolescent Chemotherapy: Potential for Epigenetic Inheritance. Plos One. 2017;12(2).
352. Fukuda K, Inoguchi Y, Ichiyanagi K, Ichiyanagi T, Go Y, Nagano M, et al. Evolution of the sperm methylome of primates is associated with retrotransposon insertions and genome instability. Human Molecular Genetics. 2017;26(18):3508-19.

353. Camprubi C, Cigliano RA, Salas-Huetos A, Garrido N, Blanco J. What the human sperm methylome tells us. Epigenomics. 2017;9(10):1299-315.
354. Chan D, McGraw S, Klein K, Wallock LM, Konermann C, Plass C, et al. Stability of the human sperm DNA methylome to folic acid fortification and short-term supplementation. Human Reproduction. 2017;32(2):272-83.
355. Jenkins TG, James ER, Alonso DF, Hoidal JR, Murphy PJ, Hotaling JM, et al. Cigarette smoking significantly alters sperm DNA methylation patterns. Andrology. 2017;5(6):1089-99.
356. Cheng YS, Lu CW, Lin TY, Lin PY, Lin YM. Causes and Clinical Features of Infertile Men With Nonobstructive Azoospermia and Histopathologic Diagnosis of Hypospermatogenesis. Urology. 2017;105:62-8.
357. Laqqan M, Tierling S, Alkhaled Y, Lo Porto C, Solomayer EF, Hammadeh M. Spermatozoa from males with reduced fecundity exhibit differential DNA methylation patterns. Andrology. 2017;5(5):971-8.
358. Laqqan M, Tierling S, Alkhaled Y, LoPorto C, Hammadeh ME. Alterations in sperm DNA methylation patterns of oligospermic males. Reproductive Biology. 2017;17(4):396-400.
359. Denomme MM, McCallie BR, Parks JC, Schoolcraft WB, Katz-Jaffe MG. Alterations in the sperm histone-retained epigenome are associated with unexplained male factor infertility and poor blastocyst development in donor oocyte IVF cycles. Human Reproduction. 2017;32(12):2443-55.
360. Laqqan M, Tierling S, Alkhaled Y, Lo Porto C, Solomayer EF, Hammadeh ME. Aberrant DNA methylation patterns of human spermatozoa in current smoker males. Reproductive Toxicology. 2017;71:126-33.
361. Pilsner JR, Shershebnev A, Medvedeva YA, Suvorov A, Wu HT, Goltsov A, et al. Peripubertal serum dioxin concentrations and subsequent sperm methylome profiles of young Russian adults. Reproductive Toxicology. 2018;78:40-9.
362. Murphy SK, Itchon-Ramos N, Visco Z, Huang ZQ, Grenier C, Schrott R, et al. Cannabinoid exposure and altered DNA methylation in rat and human sperm. Epigenetics. 2018;13(12):1208-21.
363. Jenkins TG, Aston KI, Carrell DT. Sperm epigenetics and aging. Translational Andrology and Urology. 2018;7:S328-S35.
364. Al Khaled Y, Tierling S, Laqqan M, Lo Porto C, Hammadeh ME. Cigarette smoking induces only marginal changes in sperm DNA

methylation levels of patients undergoing intracytoplasmic sperm injection treatment. Andrologia. 2018;50(1).
365. Viana J, Hannon E, Dempster E, Pidsley R, Macdonald R, Knox O, et al. Schizophrenia-associated methylomic variation: molecular signatures of disease and polygenic risk burden across multiple brain regions. Human Molecular Genetics. 2017;26(1):210-25.
366. Mansell G, Gorrie-Stone TJ, Bao YC, Kumari M, Schalkwyk LS, Mill J, et al. Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. Bmc Genomics. 2019;20.
367. Gratten J, Wray NR, Peyrot WJ, McGrath JJ, Visscher PM, Goddard ME. Risk of psychiatric illness from advanced paternal age is not predominantly from de novo mutations. Nature Genetics. 2016;48(7):718-+.
368. Hannon E, Lunnon K, Schalkwyk L, Mill J. Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. Epigenetics. 2015;10(11):1024-32.
369. Filzmoser P, Maronna R, Werner M. Outlier Identification in High Dimensions 2007 [Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.332.7176&rep=rep1&type=pdf.
370. Price EM, Cotton AM, Lam LL, Farre P, Emberly E, Brown CJ, et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics & Chromatin. 2013;6.
371. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8(2):203-9.
372. Jenkins TG, Aston KI, Cairns B, Smith A, Carrell DT. Paternal germ line aging: DNA methylation age prediction from human sperm. Bmc Genomics. 2018;19.
373. Schneider E, Dittrich M, Bock J, Nanda I, Muller T, Seidmann L, et al. CpG sites with continuously increasing or decreasing methylation from early to late human fetal brain development. Gene. 2016;592(1):110-8.
374. NCBI. dbSNP Human Build 151 database 2019 [Available from: https://www.ncbi.nlm.nih.gov/snp/.

375. Brandeis M, Frank D, Keshet I, Siegfried Z, Mendelsohn M, Nemes A, et al. Sp1 elements protect a CpG island from de novo methylation. Nature. 1994;371(6496):435-8.
376. Ritchie ME, Phipson B, Wu D, Hu YF, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7).
377. Kirchner H, Sinha I, Gao H, Ruby MA, Schonke M, Lindvall JM, et al. Altered DNA methylation of glycolytic and lipogenic genes in liver from obese and type 2 diabetic patients. Molecular Metabolism. 2016;5(3):171-83.
378. Spiers H, Hannon E, Schalkwyk LC, Smith R, Wong CCY, O'Donovan MC, et al. Methylomic trajectories across human fetal brain development. Genome Research. 2015;25(3):338-52.
379. Hodges E, Molaro A, Dos Santos CO, Thekkat P, Song Q, Uren PJ, et al. Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment. Molecular Cell. 2011;44(1):17-28.
380. Field AE, Robertson NA, Wang T, Havas A, Ideker T, Adams PD. DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. Molecular Cell. 2018;71(6):882-95.
381. Horvath S, Erhart W, Brosch M, Ammerpohl O, von Schonfels W, Ahrens M, et al. Obesity accelerates epigenetic aging of human liver. Proceedings of the National Academy of Sciences of the United States of America. 2014;111(43):15538-43.
382. Bell CG. The Epigenomic Analysis of Human Obesity. Obesity. 2017;25(9):1471-81.
383. Masser DR, Stanford DR, Freeman WM. Targeted DNA Methylation Analysis by Next-generation Sequencing. Jove-Journal of Visualized Experiments. 2015(96).
384. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research. 2010;38(6):1767-71.
385. NCBI. GRCh38.p13 Reference Genome: NCBI project PRJNA31257; 2019 [Available from: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39/.
386. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27(11):1571-2.

387. Bioinformatics B. Bismark Sequence aligner. Cambridge, UK: Babraham Institute; 2019.
388. Bioinformatics B. FastQC Cambridge, UK: Babraham Institute; 2019 [cited 2019 2019]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
389. Illumina. Quality Scores for Next-Generation Sequencing 2011 [Available from: https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf.
390. Du P, Zhang XA, Huang CC, Jafari N, Kibbe WA, Hou LF, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. Bmc Bioinformatics. 2010;11.
391. Li S, Wong EM, Bui M, Nguyen TL, Joo JHE, Stone J, et al. Inference about causation between body mass index and DNA methylation in blood from a twin family study. International Journal of Obesity. 2019;43(2):243-52.
392. Farag AK, Roh EJ. Death-associated protein kinase (DAPK) family modulators: Current and future therapeutic outcomes. Medicinal Research Reviews. 2019;39(1):349-85.
393. Geering B, Stoeckle C, Rozman S, Oberson K, Benarafa C, Simon HU. DAPK2 positively regulates motility of neutrophils and eosinophils in response to intermediary chemoattractants. Journal of Leukocyte Biology. 2014;95(2):293-303.
394. Stolarczyk E. Adipose tissue inflammation in obesity: a metabolic or immune response? Current Opinion in Pharmacology. 2017;37:35-40.
395. Stampolidis P, Ullrich A, Lacobelli S. LGALS3BP, lectin galactoside-binding soluble 3 binding protein, promotes oncogenic cellular events impeded by antibody intervention. Oncogene. 2015;34(1):39-52.
396. Niu L, Geyer PE, Albrechtsen NJW, Gluud LL, Santos A, Doll S, et al. Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease. Molecular Systems Biology. 2019;15(3).
397. Syvanen AC. Toward genome-wide SNP genotyping. Nature Genetics. 2005;37:S5-S10.
398. Tran H, Wu XW, Tithi S, Sun MA, Xie HH, Zhang LQ. A Bayesian Assignment Method for Ambiguous Bisulfite Short Reads. Plos One. 2016;11(3).
399. Chatterjee A, Stockwell PA, Rodger EJ, Morison IM. Comparison of alignment software for genome-wide bisulphite sequence data. Nucleic Acids Research. 2012;40(10).

400. Lappalainen T, Greally JM. Associating cellular epigenetic models with human phenotypes. Nature Reviews Genetics. 2017;18(7):441-51.
401. Magnus P, Birke C, Vejrup K, Haugan A, Alsaker E, Daltveit AK, et al. Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). International Journal of Epidemiology. 2016;45(2):382-8.
402. Stephenson J, Heslehurst N, Hall J, Schoenaker D, Hutchinson J, Cade JE, et al. Before the beginning: nutrition and lifestyle in the preconception period and its importance for future health. Lancet. 2018;391(10132):1830-41.
403. Milner J, Arezina J. The accuracy of ultrasound estimation of fetal weight in comparison to birth weight: A systematic review. Ultrasound. 2018;26(1):32-41.
404. Abbasi M, Smith AD, Swaminathan H, Sangngern P, Douglas A, Horsager A, et al. Establishing a stable, repeatable platform for measuring changes in sperm DNA methylation. Clinical Epigenetics. 2018;10.
405. Martinson ML, Reichman NE. Socioeconomic Inequalities in Low Birth Weight in the United States, the United Kingdom, Canada, and Australia. American Journal of Public Health. 2016;106(4):748-54.
406. Spencer N, Bambang S, Logan S, Gill L. Socioeconomic status and birth weight: comparison of an area-based measure with the Registrar General's social class. Journal of Epidemiology and Community Health. 1999;53(8):495-8.
407. Weightman AL, Morgan HE, Shepherd MA, Kitcher H, Roberts C, Dunstan FD. Social inequality and infant health in the UK: systematic review and meta-analyses. Bmj Open. 2012;2(3).
408. Connolly V, Unwin N, Sherriff P, Bilous R, Kelly W. Diabetes prevalence and socioeconomic status: a population based study showing increased prevalence of type 2 diabetes mellitus in deprived areas. Journal of Epidemiology and Community Health. 2000;54(3):173-7.
409. Stringhini S, Zaninotto P, Kumari M, Kivimaki M, Batty GD. Lifecourse socioeconomic status and type 2 diabetes: the role of chronic inflammation in the English Longitudinal Study of Ageing. Scientific Reports. 2016;6.
410. Bann D, Johnson W, Li L, Kuh D, Hardy R. Socioeconomic Inequalities in Body Mass Index across Adulthood: Coordinated Analyses of Individual Participant Data from Three British Birth Cohort Studies Initiated in 1946, 1958 and 1970. Plos Medicine. 2017;14(1).

411. Hardcastle TJ, Lewsey MG. Mobile small RNAs and their role in regulating cytosine methylation of DNA. Rna Biology. 2016;13(11):1060-7.
412. Birney E, Smith GD, Greally JM. Epigenome-wide Association Studies and the Interpretation of Disease -Omics. Plos Genetics. 2016;12(6).
413. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. Nature Reviews Genetics. 2018;19(3):129-47.
414. Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. Nature Reviews Genetics. 2013;14(8):585-94.
415. Relton CL, Smith GD. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. International Journal of Epidemiology. 2012;41(1):161-76.
416. Valezi AC, Menezes MD, Mail J. Weight Loss Outcome After Roux-en-Y Gastric Bypass: 10 Years of Follow-up. Obesity Surgery. 2013;23(8):1290-3.
417. McGrath JJ, Petersen L, Agerbo E, Mors O, Mortensen PB, Pedersen CB. A Comprehensive Assessment of Parental Age and Psychiatric Disorders. Jama Psychiatry. 2014;71(3):301-9.
418. Soubry A, Schildkraut JM, Murtha A, Wang F, Huang ZQ, Bernal A, et al. Paternal obesity is associated with IGF2 hypomethylation in newborns: results from a Newborn Epigenetics Study (NEST) cohort. Bmc Medicine. 2013;11:10.
419. Warrington NM, Beaumont RN, Horikoshi M, Day FR, Helgeland O, Laurin C, et al. Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. Nature Genetics. 2019;51(5):804-+.

Appendices

Appendix 1. Questionnaire for Participating Fathers

University College London Hospitals NHS
NHS Foundation Trust

The Dad's Health Study

(A Study to Investigate the Influence of Paternal Health on his Baby's Birthweight)

QUESTIONNAIRE FOR FATHERS V2 7/2/2016

This questionnaire asks you some questions about your health and background.

All answers you give are confidential and will only be seen by the research team.

We would be grateful if you would help us by answering all the questions.

If you have any queries about any of the questions or would like some help in completing this questionnaire, please contact Dr Fredrika Asenius on 07926668506 or karin.asenius.12@ucl.ac.uk

Thank you very much for your help.

Name

_____

Date of Birth

_____

Participant Identification Number

_____

Date today

_____

Length of gestation

_____

Details (in order order to be contact you with relevant results)

Mobile/home telephone

_____

Home Address

_____

GP Details

_____

Partner's name

_____

1.	How old are you?			
2.	Do you know what your weight is approximately?			
3.	Do you know what your height is approximately?			
4.	Is this your first child with your current partner? (Circle as appropriate)	Yes	No	
	If no, how many children do you have with your current partner?			
5.	Do you have any medical problems?	Yes	No	
	If yes:			
	Diabetes?	Yes	No	
	Requiring insulin?	Yes	No	
	Heart disease?	Yes	No	
	Kidney disease?	Yes	No	
	Haematological problems?	Yes	No	
	Other (please specify)			
6.	Are you taking any medications?	Yes	No	
	If yes, which?			
7.	Do you know how much you weighed when born?			
8.	Were you called a 'small' baby?	Yes	No	DK

9.	Were you born early?	Yes	No	DK		
	If so, at how many weeks?					
10.	Were you a twin?	Yes	No	DK		
11.	Do you recall if your mother had high blood pressure in her pregnancy /pregnancies?	Yes	No	DK		
12.	Did / do either of your parents have significant health problems?	Yes	No			
	If yes, can you tell us whom and what?					
13.	How long did you have unprotected sex or non-barrier contraception, before conception of this pregnancy? (Circle as appropriate)	<3 months	3-6 months	6-12 months	1-3 years	>3 years
14.	Was this a natural or IVF pregnancy?					
15.	Have you ever smoked?	Yes	No			
	If previously, how long ago and how many?	0-5	5-10	10-20	20+	
16.	Do you drink alcohol?	Yes	No			
	If yes, approximately how many glasses*/ week * By glass we mean a pub measure of spirits, half a pint of lager or cider, a wine glass of wine, etc	None	1-6 glasses	7-20 glasses	21+ glasses	
17.	If this is not your first child with this partner, has there been any significant change in your lifestyle between children i.e. smoking, exercise, development of disease or significant weight gain or loss?					

18.	Anything else you would like to share that may be relevant?	

Thank You

Appendix 2. Questionnaire for Participating Mothers

University College London Hospitals NHS
NHS Foundation Trust

The Dad's Health Study

(A Study to Investigate the Influence of Paternal Health on his Baby's Birthweight)

QUESTIONNAIRE FOR MOTHERS V2 21/2/2016

This questionnaire asks you some questions about your health and background.

All answers you give are confidential and will only be seen by the research team.

We would be grateful if you would help us by answering all the questions.

If you have any queries about any of the questions or would like some help in completing this questionnaire, please contact Dr Fredrika Asenius on 07926668506 or karin.asenius.12@ucl.ac.uk

Thank you very much for your help

Name

_____

Date of Birth

_____

Participant Identification Number

_____

Date today

_____

Length of gestation (how many weeks pregnant you are) and Estimated Due Date (EDD)

_____

Details (in order to contact you with relevant results)

Mobile/home telephone

_____

Home Address

_____

GP Details

_____

Partner's name

_____

1.	How old are you?		
2.	What was your pre-pregnancy weight approximately?		
3.	What is your height approximately?		
4.	Is this your first child with your current partner? (Circle as appropriate)	Yes	No
	If no, how many children do you have with your current partner?		
	What weights were these children at birth?		
	Were any of your children born early (please give details)?		
5.	Have you been admitted to hospital in this pregnancy?	Yes	No
	If yes, what was the reason for admission?		
6.	Do you have any medical problems?	Yes	No
	If yes:		
	Diabetes?	Yes	No
	Heart disease?	Yes	No
	Kidney disease?	Yes	No
	Haematological problems?	Yes	No
	Other (please specify)		

7.	Are you taking any medications?	Yes			No	
	If yes, which?					
8.	What was your weight at birth?					
9.	Were you called a 'small' baby?	Yes		No		DK
10.	Were you born early?	Yes		No		DK
11.	Were you a twin?	Yes		No		DK
12.	Did your mother have high blood pressure in her pregnancy /pregnancies?	Yes		No		DK
13.	Did / do either of your parents have significant health problems?	Yes			No	
	If yes, can you tell us whom and what?					
14.	How long did you have unprotected sex or non-barrier contraception, before conception of this pregnancy? (Circle as appropriate)	<3 months	3-6 months	6-12 months	1-3 years	>3 years
15.	Was this a natural or IVF pregnancy?					
16.	Have you ever smoked?	Yes			No	
	If previously, how long ago?	0-5y	5-10y	10-20y	20+y	
	How many cigarettes do or did you smoke per day?					
	If yes, how many/day during this pregnancy?	0-5	5-10	10-20	20+	
17.	How much alcohol have you drunk at the following times?					
	In the month before pregnancy how many glasses*/ week * By glass we mean a pub measure of spirits, half a pint of lager or cider, a wine	None	1-6 glasses	7-20 glasses	21+ glasses	

	glass of wine, etc				
	In first three months of pregnancy?	None	1-6 glasses	7-20 glasses	20+ glasses
18.	If this is not your first child with this partner, has there been any significant change in your lifestyle between children i.e. smoking, exercise, development of disease or significant weight gain or loss?				
19.	Anything else you would like to share that may be relevant?				

Thank You

Appendix 3. List of the 192 primers included in the study of obesity associated CpG methylation in blood

Primer Number	Probe	Marker	Size	Class (1-5)	Dimer (+/-)
1	cg01024458	Blood.B	100	5	-
2	cg04162316	Blood.CD4	232	5	+
3	cg09315878	BMI.EWAS	323	3	-
4	cg09554443	BMI.EWAS	106	1	-
5	cg09491962	BMI.EWAS	271	5	-
6	cg26673975	Blood.Gran	348	4	+
7	cg16395997	BMI.EWAS	196	4	+
8	cg09935388	BMI.EWAS	325	3	-
9	cg26257082	BMI.EWAS	222	3	-
10	cg07504977	BMI.EWAS	171	5	-
11	cg25131632	Blood.B	100	5	-
12	cg10837404	Blood.CD4	286	3	-
13	cg18990407	Blood.Mono	326	3	+
14	cg24145109	BMI.EWAS	254	5	-
15	cg03725309	BMI.EWAS	206	5	-
16	cg23172671	BMI.EWAS	271	5	+
17	cg00431050	BMI.EWAS	285	5	-
18	cg13738327	Blood.B	333	5	-
19	cg26166854	Blood.CD4	350	3	+
20	cg26961332	Blood.Gran	323	3	-
21	cg23244761	Blood.Mono	273	3	-
22	cg09032544	Blood.CD8	317	5	-
23	cg11130778	Blood.Gran	350	5	-
24	cg03538296	Blood.NK	306	5	-
25	cg11673687	BMI.EWAS	350	3	-
26	cg01455178	BMI.EWAS	308	3	-
27	cg27106643	Blood.B	112	3	-
28	cg11067179	Blood.CD8	228	3	-
29	cg15013257	Blood.Gran	120	3	-
30	cg12484113	BMI.EWAS	201	2	-
31	cg13997435	BMI.EWAS	260	4	-
32	cg10717869	BMI.EWAS	161	3	-
33	cg17768768	Blood.B	289	5	+
34	cg06164961	Blood.CD8	272	5	+
35	cg17822325	BMI.EWAS	115	3	-
36	cg15323828	BMI.EWAS	288	3	+

37	cg00244001	BMI.EWAS	350	3	-
38	cg13828440	Blood.NK	147	5	+
39	cg12593793	BMI.EWAS	119	2	-
40	cg01101459	BMI.EWAS	284	3	+
41	cg13400249	Blood.CD4	350	5	+
42	cg13430807	Blood.Mono	163	5	-
43	cg00851028	BMI.EWAS	338	5	-
44	cg18174654	Blood.CD8	121	4	-
45	cg00701951	Blood.Mono	254	4	-
46	cg25217710	BMI.EWAS	298	3	+
47	cg02079413	BMI.EWAS	274	5	+
48	cg07679948	Blood.CD4	348	4	-
49	cg04468741	Blood.Mono	231	5	-
50	cg10092518	BMI.EWAS	275	4	-
51	cg11986385	BMI.EWAS	282	4	-
52	cg08777095	Blood.CD8	133	4	-
53	cg16636767	Blood.Mono	320	3	+
54	cg25001190	BMI.EWAS	221	5	-
55	cg04869770	BMI.EWAS	332	4	-
56	cg07136133	BMI.EWAS	287	4	-
57	cg17260706	BMI.EWAS	344	4	-
58	cg15159104	BMI.EWAS	350	5	-
59	cg02426464	BMI.EWAS	350	5	-
60	cg24824917	BMI.EWAS	230	3	-
61	cg00417304	BMI.EWAS	123	5	-
62	cg26894079	BMI.EWAS	298	4	-
63	cg21670987	BMI.EWAS	121	3	-
64	cg01798813	BMI.EWAS	212	4	-
65	cg24174557	BMI.EWAS	277	5	-
66	cg11376147	BMI.EWAS	207	4	-
67	cg07217499	BMI.EWAS	317	5	-
68	cg03508235	BMI.EWAS	281	5	-
69	cg07037944	BMI.EWAS	158	2	-
70	cg08877257	BMI.EWAS	182	4	-
71	cg19217955	BMI.EWAS	105	3	+
72	cg13243168	BMI.EWAS	130	5	-
73	cg12917475	BMI.EWAS	234	3	-
74	cg02119938	BMI.EWAS	322	1	-
75	cg22695339	BMI.EWAS	282	3	+
76	cg18772573	BMI.EWAS	282	4	+
77	cg03433986	BMI.EWAS	168	5	-
78	cg03523676	BMI.EWAS	347	5	-

79	cg07728579	BMI.EWAS	245	4	-
80	cg14020176	BMI.EWAS	223	5	-
81	cg00994936	BMI.EWAS	193	3	-
82	cg27117792	BMI.EWAS	297	5	-
83	cg26357885	BMI.EWAS	191	5	-
84	cg20507228	BMI.EWAS	215	5	-
85	cg11024682	BMI.EWAS	291	1	+
86	cg10508317	BMI.EWAS	280	5	-
87	cg11152384	BMI.EWAS	292	4	-
88	cg27614723	BMI.EWAS	124	2	-
89	cg00863378	BMI.EWAS	350	2	-
90	cg16611584	BMI.EWAS	261	3	-
91	cg22950899	BMI.EWAS	215	3	+
92	cg11660018	BMI.EWAS	316	4	-
93	cg00973118	BMI.EWAS	267	4	+
94	cg24457403	BMI.EWAS	310	4	-
95	cg26651978	BMI.EWAS	219	5	-
96	cg19574327	BMI.EWAS	226	5	-
97	cg25096107	BMI.EWAS	153	3	-
98	cg08857797	BMI.EWAS	303	2	-
99	cg02008402	BMI.EWAS	267	5	+
100	cg10734665	BMI.EWAS	209	4	-
101	cg03078551	BMI.EWAS	249	5	-
102	cg09777883	BMI.EWAS	349	4	-
103	cg19750657	BMI.EWAS	164	5	+
104	cg01419914	BMI.EWAS	136	5	-
105	cg07814318	BMI.EWAS	339	5	-
106	cg09109383	BMI.EWAS	206	4	+
107	cg27050612	BMI.EWAS	339	4	+
108	cg20981127	BMI.EWAS	348	5	-
109	cg06876354	BMI.EWAS	218	4	-
110	cg00916899	BMI.EWAS	213	5	-
111	cg04924511	BMI.EWAS	288	5	-
112	cg19266387	BMI.EWAS	146	5	-
113	cg15442888	BMI.EWAS	244	5	-
114	cg13305415	BMI.EWAS	233	5	-
115	cg15835542	BMI.EWAS	229	5	-
116	cg04557677	BMI.EWAS	188	5	-
117	cg15357118	BMI.EWAS	287	5	-
118	cg18217136	BMI.EWAS	241	4	-
119	cg13010621	BMI.EWAS	159	5	-
120	cg03327570	BMI.EWAS	232	5	-

121	cg05628049	BMI.EWAS	227	5	-
122	cg23576855	BMI.EWAS	190	4	+
123	cg16721489	BMI.EWAS	189	4	+
124	cg08309687	BMI.EWAS	349	5	-
125	cg02286155	BMI.EWAS	311	1	-
126	cg17178175	BMI.EWAS	236	5	-
127	cg01881899	BMI.EWAS	268	1	+
128	cg04816311	BMI.EWAS	160	4	+
129	cg02560388	BMI.EWAS	345	5	+
130	cg00108715	BMI.EWAS	334	5	-
131	cg22143698	BMI.EWAS	276	5	-
132	cg09222732	BMI.EWAS	276	5	-
133	cg23647610	BMI.EWAS	344	5	-
134	cg08548559	BMI.EWAS	350	2	-
135	cg10179300	BMI.EWAS	331	1	-
136	cg09956615	BMI.EWAS	158	4	-
137	cg19017142	BMI.EWAS	132	4	-
138	cg27115863	BMI.EWAS	269	2	-
139	cg13084458	BMI.EWAS	259	5	-
140	cg08215255	BMI.EWAS	210	5	-
141	cg24776142	BMI.EWAS	299	2	+
142	cg00634542	BMI.EWAS	196	4	-
143	cg07730360	BMI.EWAS	255	5	-
144	cg06690548	BMI.EWAS	336	1	+
145	cg03957124	BMI.EWAS	257	5	-
146	cg26804423	BMI.EWAS	262	4	-
147	cg23417875	BMI.EWAS	218	5	-
148	cg03318904	BMI.EWAS	168	4	-
149	cg01671681	BMI.EWAS	122	5	-
150	cg01300684	BMI.EWAS	226	4	-
151	cg25570328	BMI.EWAS	281	5	+
152	cg09349128	BMI.EWAS	269	1	+
153	cg00673344	BMI.EWAS	265	5	+
154	cg00850073	BMI.EWAS	329	4	-
155	cg00585790	BMI.EWAS	305	4	+
156	cg26361535	BMI.EWAS	277	1	-
157	cg02650017	CRP/BMI	250	4	+
158	cg08118908	Smoking	299	5	+
159	cg25197194	Smoking/BMI	280	5	-
160	cg01677628	Obesity SNP associated	180	4	+
161	cg26545918	Obesity SNP associated	277	5	-
162	cg10062919	Smoking	344	4	-

163	cg09363892	Obesity SNP associated	153	5	-
164	cg10639395	Obesity SNP associated	304	5	-
165	cg00834536	Obesity SNP associated	267	5	-
166	cg13591783	BMI.EWAS	316	4	-
167	cg22304262	CRP/BMI	286	5	-
168	cg27269962	BMI.EWAS	213	5	-
169	cg13840239	BMI.EWAS	163	5	-
170	cg12992827	CRP/BMI	260	5	+
171	cg00073090	Smoking	159	5	-
172	cg01844514	BMI.EWAS	317	4	-
173	cg14264316	BMI.EWAS	199	4	-
174	cg17560136	BMI.EWAS	191	4	-
175	cg19406367	Smoking	320	4	-
176	cg02571142	BMI.EWAS	330	5	-
177	cg19589396	BMI.EWAS	127	5	-
178	cg14476101	CRP	349	4	-
179	cg11028075	Smoking	290	4	-
180	cg26077378	Smoking	295	5	+
181	cg22864340	Obesity SNP associated	257	4	+
182	cg00489954	Obesity SNP associated	240	4	+
183	cg23078228	Obesity SNP associated	219	4	-
184	cg07960624	BMI.EWAS	181	4	-
185	cg12054453	CRP	186	4	+
186	cg19821297	CRP	261	4	+
187	cg11700584	Smoking	291	4	+
188	cg26952928	BMI.EWAS	308	4	+
189	cg17592360	Obesity SNP associated	127	5	-
190	cg25392060	BMI.EWAS	195	1	-
191	cg15497724	Obesity SNP associated	177	4	-
192	cg05149343	Obesity SNP associated	253	5	+

Appendix 3 Table 1. List of primers included in the profiling of obesity associated CpG methylation in blood.

'Marker' refers to the reason of primer inclusion where BMI.EWAS = identified from EWASs of BMI, smoking = control probes for smoking, obesity SNP associated = sites previously identified as influenced by obesity associated SNPs, and CRP = control probes for inflammation. Blood.Mono, Blood.CD4, Blood.Gran, Blood.B and Blood.NK are all control probes for blood cell composition.

Appendix 4. FDR corrected p-values for the t-test of DNA methylation differences between the lean and obese participants in the study of obesity associated CpG methylation in blood

Probe	FDR corrected p
cg07037944	0.20124822
cg26651978	0.35251394
cg12917475	0.58270552
cg07960624	0.58270552
cg03078551	0.58270552
cg14264316	0.58270552
cg09349128	0.58270552
cg15442888	0.58270552
cg00108715	0.58270552
cg08857797	0.58270552
cg03957124	0.58270552
cg15357118	0.58270552
cg10508317	0.58270552
cg07504977	0.58270552
cg13997435	0.58270552
cg00916899	0.58270552
cg09363892	0.58270552
cg11376147	0.58270552
cg05628049	0.58270552
cg20507228	0.58270552
cg00634542	0.58270552
cg00431050	0.58270552
cg06876354	0.69132912
cg25570328	0.70910387
cg03327570	0.73125804
cg10717869	0.73125804
cg13084458	0.74244415
cg01677628	0.74244415
cg01881899	0.74244415
cg11152384	0.74244415
cg25217710	0.74244415
cg01798813	0.74244415
cg26257082	0.74244415
cg16611584	0.74244415
cg16721489	0.74244415
cg25096107	0.74244415
cg04924511	0.74244415
cg00994936	0.74244415
cg16395997	0.78121714
cg02008402	0.78121714
cg23417875	0.78121714
cg17560136	0.78121714
cg19750657	0.78121714
cg27269962	0.78121714
cg08877257	0.78121714
cg11660018	0.78121714
cg00834536	0.78121714
cg18217136	0.78121714
cg04557677	0.78121714
cg09956615	0.78121714
cg10734665	0.78530064
cg24824917	0.83358549
cg15497724	0.83574834
cg05149343	0.83704821
cg09109383	0.83704821
cg23576855	0.93396605
cg23172671	0.95211277
cg07728579	0.98210114
cg13010621	0.98378658
cg13840239	0.98378658
cg14020176	0.98378658
cg22143698	0.98378658
cg00489954	0.98378658
cg03433986	0.98378658
cg22950899	0.99158091
cg04816311	0.99158091
cg09222732	0.99275048
cg01101459	0.99275048
cg26357885	0.99275048
cg24145109	0.99275048

Appendix 5. Example of a FastQC quality control report of the Illumina MiSeq data in the study of obesity associated CpG methylation in blood

FastQC Report

Tue 2 Jul 2019
1138_S127_L001_R1_001.fastq.gz

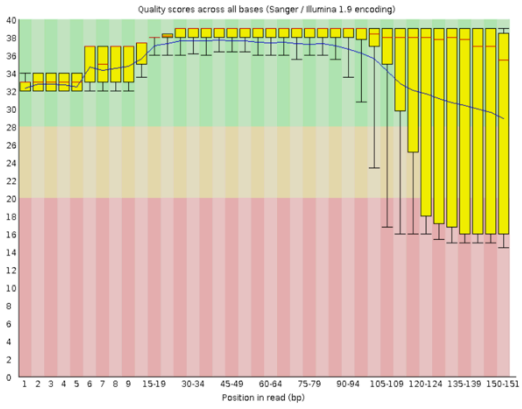
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

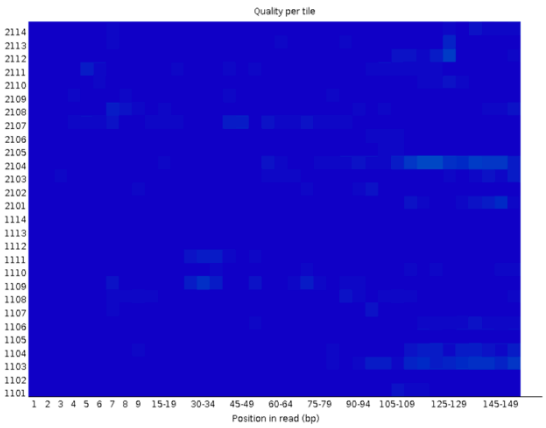
Basic Statistics

Measure	Value
Filename	1138_S127_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	58518
Sequences flagged as poor quality	0
Sequence length	151
%GC	38

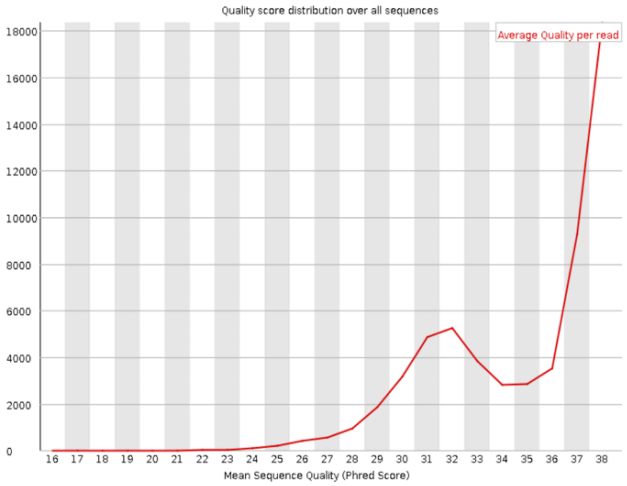
Per base sequence quality



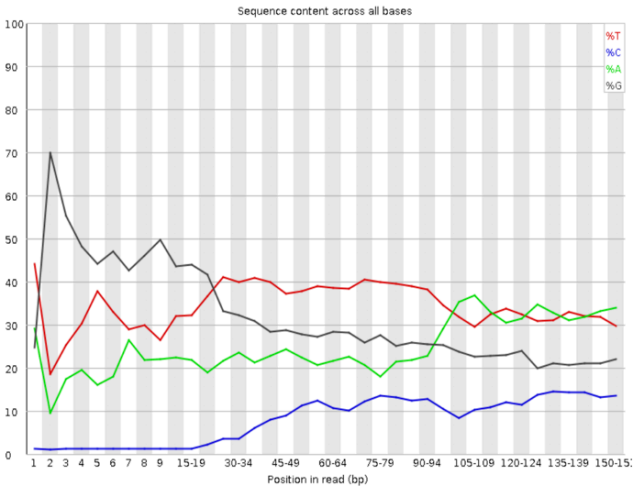
Per tile sequence quality



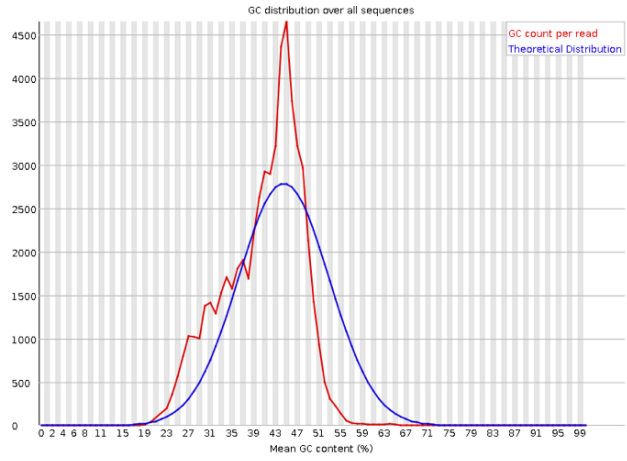
Per sequence quality scores



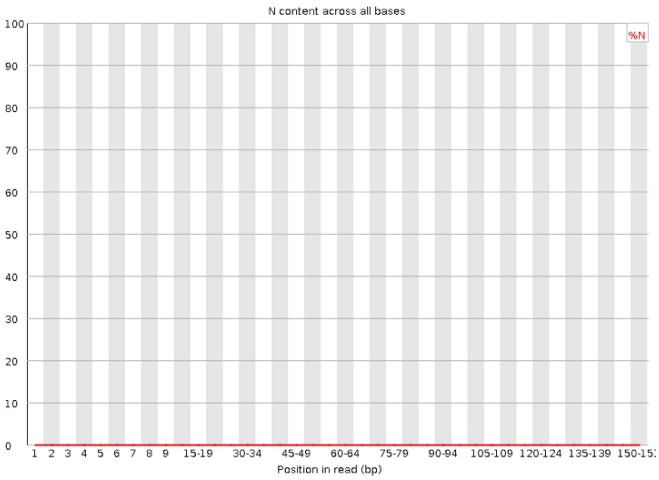
Per base sequence content



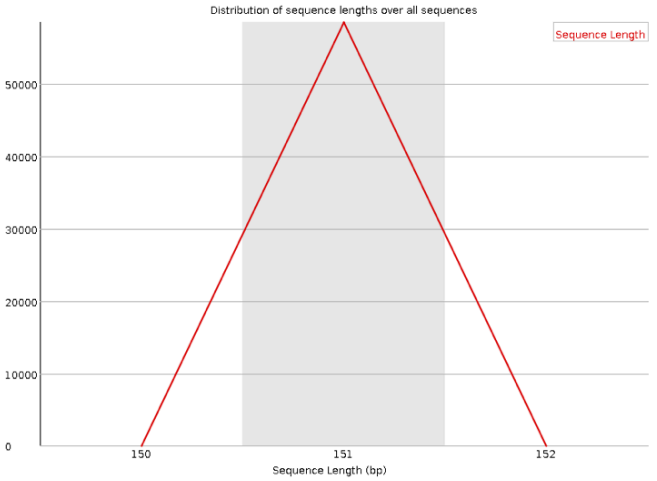
Per sequence GC content



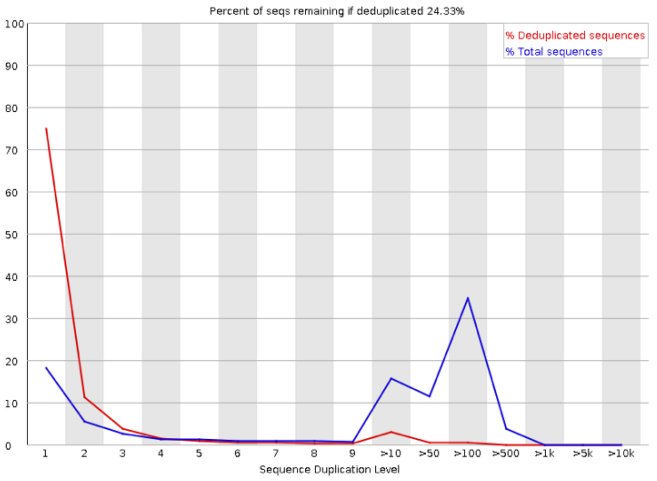
Per base N content



Sequence Length Distribution



Sequence Duplication Levels

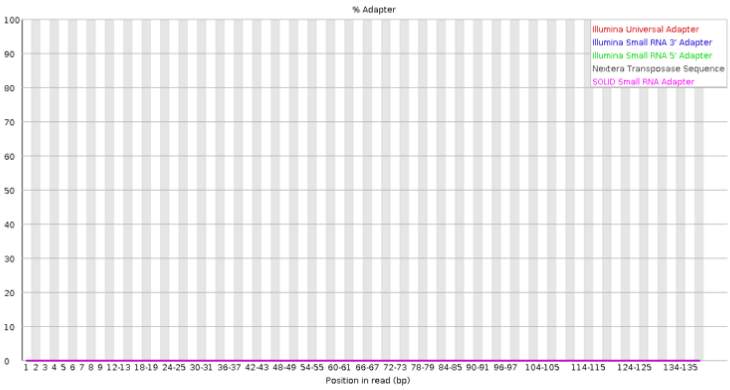


Overrepresented sequences

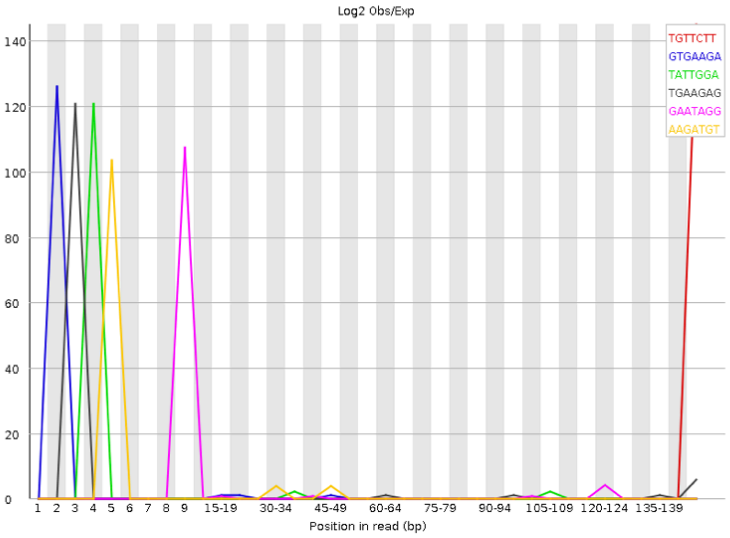
Sequence	Count	Percentage	Possible Source
AGGGGGAAGAATAGGATGGTAGCGTTTTTATTTGGGGTCGTTTTAGTTTC	752	1.2850746778768927	No Hit
AGGGTAGTAGATAGTGTTTGTGTTTGGGGTAGGAATTTAGTGAGTTAGAC	547	0.9347551180833249	No Hit
TGTAGGATTGGTTTTAGGTTGGTTATTTAGTTTGTTTGAAGTTGGATAGA	540	0.9227929867733005	No Hit
TTGGGGTGGGAATTTGGAGTTTGTATTTTTATAGAATTTTTTAGGTAAGT	517	0.8834888410403637	No Hit
TGGGTGGTTTGTGGGAAAGAGAGTAGTTAGTATTTGGATTTTGTTAGTGG	462	0.7895006664616017	No Hit
TTGTGTGTGTAGAGAAGGGGAGTTGGAGGTTAGACCAAGTCTCTGCTACC	443	0.7570320243343929	No Hit
AGGAAAGGAAAGATGGGGAAGTTTAGATTATTAGGTTTTTATTAAGGAGG	442	0.7553231484329608	No Hit
TGGGTTGAGTTGTGTTTGTTGAATTGTATTTTTTTGTTTTTGTGAAGTA	432	0.7382343894186404	No Hit
GGGGATAGGAATGAGGGTGGGTGGTTTGTGTGAATAGACCAAGTCTCTGC	412	0.7040568713899997	No Hit
TTGGGGAGGAGTTTGGAAGGGTTTGTAGTTATTGGTTTTTTTTTTTTTTT	401	0.6852592364742472	No Hit
GGTTGGAAGGAGTGTTATGGGATTATTTAGGTGTAGATTGTAGAGGATGT	397	0.6784237328685191	No Hit
AGTGAAGAGAGGAGGTTATTGAGATTACGATTATCGCGTTTTGGGTGATA	379	0.64766339666427424	No Hit
TGTGTGAAATGTTATAGGGAGTAGGAGTTATATTTTGTGATTTGTAGTTA	361	0.6169042004169657	No Hit
AGGATGGAGAGGAGTGTTTTGAGATTATTTAGGTGTAGATTGTAGAGGAT	351	0.5998154414026453	No Hit
TTGTGGAGGGTGAAATGGGGTAGGGTAGGGAAGTTGGTAGACCAAGTCTC	342	0.584435558289757	No Hit
TGGGGGTTTTTGTGAAGGTGAGTATAGTAGATAGGGTAATTTGTTATAGA	331	0.5656379233740046	No Hit
TTGTTGTGGGAATGTTTAAAGGTTAGTTTATTATAGATTAGGTGTTTTTA	323	0.5519669161625482	No Hit
TGGGGATGGGAGTTTGGTTTAGTTGGTTTGAGGTTTTTAGACCAAGTCTC	321	0.5485491643596841	No Hit
TTGAGTATTTGGGTGGTTGGAAATTTAAGAACGGTGTTTAGTTCGGTTTT	311	0.5314604053453638	No Hit
GGGTATTGGAGGGAAGGGGATATAGTCGTTTTGTTGAGGGTTAATTTTG	309	0.5280426535424997	No Hit
GGGAGAGAAAGGTTTTATGGGTGGGGGTTAAGGGTAGGAGTTTTGGAGTT	304	0.5194982740353395	No Hit
TGGAGAAGAAAGTTGTTGAGAGGAGTAGGTTATGTTTGTTTTATTGTTT	296	0.5058272668238832	No Hit
TTGGGGATTTAGGGAGTTAGTTTTTAGTTTTGGTTTACGAATTCGTTTTA	295	0.5041183909224511	No Hit
AGTGTGGTAGAAGAGGAAAGGTAGGGTTAGGTAATTGGTATTTAGACCAA	290	0.4955740141529106	No Hit
TGGGTTGGTGTTTGGGGAATAGTGTGTTTTTGTTTAGCGTAGTATAGATG	285	0.4870296319081308	No Hit
GTGGTTTGGGGAAATAGGGTTTATTTAAGGTTTGGTTTAGGAAAGGTTTT	273	0.4665231210909464	No Hit
GAAAGGAAGTGGGTGTGGGTTTGGTTTGTGTGTTTATTTAAGTTAGACCA	272	0.4648142451895143 6	No Hit
GGTTGTGGGAATGGGGAGTTTAGGGAAGGTTTTTAGAGTTGGGTAAGTGG	266	0.4545609897809221	No Hit
TGGGTTGGGTTTGTTAAGTTGATTAGTGTGAATTGGTTGATTGAGGAAGG	264	0.4511432397805805	No Hit
AAGAAAATGAAAATTGAGGTGTTGAAAAAAGTTTTTAAGGTTAACGTGAA	260	0.4443077343723299	No Hit
TGGGTTGGTGTTTGGGGAATAGTGTGTTTTTGTTTAGTGTAGTATAGATG	258	0.4408899825694658	No Hit
TGTAGGGGATTATAGGGTAGGGTTTACGTGTTAAGTGTGCGTTATACGTG	257	0.4391811066680337	No Hit
GTGTTTATTTTGGAGGGGGAAGGTTAGACCAAGTCTCTGCTACCGTATAG	255	0.43576335486516965	No Hit
TGGGAGGTAGAGGTTGAGGTAGTGGTGGTTGGGATGGTTAGGAGAAAAGG	249	0.4255100994565775	No Hit
TGTTGTTGGGGTTGGAGAGGGGGTGGATGGAGGATTGTATGTTGTGAGGT	245	0.4186745958508493	No Hit
GGATTGGTGGGGGTTTGGAATTAGTTGTTTAAGGAAAACGATTTTTTTAA	240	0.4101302163436891	No Hit
AGGAGTTGGATAAGTTTGAAGTTGTTTAGTTTGGGTTTTTTAGTAGAATT	237	0.405003588639393	No Hit
TGTGGTAGAGTTGGTGGGAGGTTGGAGGTGTTGGTGTGGTTTAAGGGTTA	235	0.40158583683652893	No Hit
AGGGTTTGGGTTATATTAGGGGTTTAGTTGATTTTAGATTTATTTGTTAA	218	0.3725349465121843	No Hit
TTGAGTATTTGGGTGGTTGGAATTTGGGTAGAAGGGGTTAGACCAAGTCT	217	0.37082607061075223	No Hit
TTGTGTGTGTGGAGAAGGGGAGTTGGAGGTTAGACCAAGTCTCTGCTACC	216	0.3691171947093202	No Hit
AGGTTTTAGGTTTTAGGGTGTGAATAGAGTATTAGACCAAGTCTCTGCTA	215	0.36740831880788816	No Hit
GGATTGGTGGGGGTTTGGAATTAGTTGTTTAAGGAAAACGATTTTTTTAA	212	0.36228169110359204	No Hit
AGGTGTGTTTTGTGTTGTTAGGAAATTGTAGGGTTTTGAGTATAAATTGA	211	0.36057281520216	No Hit
TTGAGGTTTAGGTTAGGGGATTTGGGGTTGTTGGAGGTAGAAGTAGACCA	210	0.35886393930072796	No Hit
GTTTTTGGGGTTAGGGGGAGTGGTTGTTGAGGGAAGGAGATAGACCAAGT	210	0.35886393930072796	No Hit
TGTAGGAAGTAATTTGATGGTGTTAATGATAATGGAGATTTTAGACCAAG	209	0.3571550633992959	No Hit
GGAAGGAAGTAGAGTTTGGGGGTTTAGATCGTATGTTTTAAAAGTTTTTT	206	0.35202843569499986	No Hit
ATGGGAGTGGGGGTTATGGGGGTTGTGGGAGTAGACCAAGTCTCTGCTAC	198	0.3383574284835435	No Hit
TTTTTGAGAGATGGGGTTTTGTTATTTTGTTTAGGTTAGTTTTGAATATT	197	0.33664855258211146	No Hit
GGAGGGGTAGGGTAAGGGTTAGGTGTGGGGGATTTAGGTGAGTATATAAG	196	0.3349396766806794	No Hit
TGAGTTAGGGGGTAAGGTGGTTGGAAATGTAGGGATAGACCAAGTCTCTG	196	0.3349396766806794	No Hit
TGGGAGAAGGGAAGATTGTGAGGGGTAGAATGTTTAGAATAGACCAAGTC	184	0.31443316586349496	No Hit
TGTTGTTGGGGTTGGAGAGGTTGTAGTTGGAAGATGGGTTAGACCAAGTC	182	0.31101541406063093	No Hit
AGTGGAGGATGGATATTGTTGTTATTTTTTTGTTTTGGTATTTGGGTTG	182	0.31101541406063093	No Hit
AAGTTAGGAGGTGGAAGGGAAGATTGTGAGGGGTAGAATGTTTAGAATAG	181	0.3093065381591989	No Hit
TGTAGGAAGTAATTTGAGGGTGTTAATGATAATGGAGATTTTAGACCAAG	179	0.3058887863563348	No Hit
GGTGGTAGTTTGGGGAGGTGGTAGTTTAGACCAAGTCTCTGCTACCGTAT	173	0.29563553094774 26	No Hit
TGGGTTGGGTTTGTTAAGTTGATTAGTGTGAATTGGTCGATCGAGGAAGG	173	0.29563553094774 26	No Hit
GGGGAGGATTTGGTTAGTGAGGGGTAGGGGAATTTATAGTAGTTTTTGTG	173	0.29563553094774 26	No Hit
AGATTGGGTGATTTTATGGTGTTTTAATGTGTGGGTTTTGATTGATTATA	171	0.2922177791448785	No Hit
AGTTTTTAAGGAAATGGAGGAGTTGTAGGGAAGAGGTGGGGTTTAGACCA	168	0.2870911514405824	No Hit
TGGGAGGTAGAGGTTGAGGTAGCGGTGGTCGGGACGGTTAGGAGAAAAGG	168	0.2870911514405824	No Hit
AAAGGATGGGTTAAAATAGTTGAGAGTGTGTTTGTTTTTTTTTGTTTTGG	165	0.2819645237362863	No Hit
GTGGAGAGAGGGTTATTGTGAGGATTTAAGAGAGTTATTAAAGGGTATTAA	161	0.2751290201305581	No Hit
TGGTTAGGGGATGGTGGGAGGTAATAGGTTATTGTAGAGTAGACCAAGTC	160	0.2734201442291261	No Hit
TGTATAGGTTTATGGGGGTGGTGGGAGTAGACCAAGTCTCTGCTACCGTA	160	0.2734201442291261	No Hit
GGGAGGAATGAGGGAATGGGGGAGGTAGTTTTGGTGGGTAGACCAAGTCT	158	0.270002392426262	No Hit
TAGTGGGAAGGTGAGGGTGGTTATAGACCAAGTCTCTGCTACCGTATAGA	157	0.26829351652482997	No Hit
TGTTTTATTGTGAGGATTTAGAGGGATATAAGTGTAGATAGTGTTGATTT	154	0.26316888882053385	No Hit
TGGTAGTTTAGGAGGTTTAGTAGGATGGGGGTTGGTGTGTAGACCAAGTC	152	0.2597491370176698	No Hit
AGAGGGGGTTGTAGGTGAGTTTTTAATGTTAGGAAATTAAGGAAGGTTTT	148	0.2529136334119416	No Hit
TGGGTAGTGGGGTAAGTTTGTTAGATTAAGGTTAAGAGTTAGAGGTTTAT	147	0.2512047575105096	No Hit
TGTTGTTGGGGTTGGAGAGGAAAGGAAGGTTTGGTAGAAGTTTTTAGAGG	145	0.24778700570764553	No Hit
GAGGTAGAGTAGAGGTTTTTGTTAGAGAGATGGGAGAAAGTTTGAGGGTT	145	0.24778700570764553	No Hit
GAGGTAGAGTAGAGGTTTTTGTTAGATTGTTTATGATGTAGGTGTAGAAT	143	0.24436925390478145	No Hit
TTTAAGGGATTTTGTTTATGGTGGTCGGTAGTTGAGGAGGTAATGTTTGT	143	0.24436925390478145	No Hit
GGAGGTTGGGGAGAGTTTGGTTAAGTTTATAGGTAGTTTTAGGATTTTTT	134	0.2289893707918930 8	No Hit
TGAGTTAGGGGGTAGGGTGGTTGGAAATGTAGGGATAGACCAAGTCTCTG	132	0.22557161898902903	No Hit
AGGTAGGGGAGGGAAAGTGGTTAGGTGATTGTAGGGGATAGACCAAGTCT	129	0.2204449912847329	No Hit
AGTGTTTTAGGGGTTGATAATTTGTTTTTAGTTTTTAGATTTATAGTTA	129	0.2204449912847329	No Hit
AGGGATTTTGTTAGAGTTGGTTTTGTTGAATTGTATTTTTTTTGTTTTTG	128	0.21873611538330084	No Hit
TGAGTTAGGGGGTAGGGTGGTTAGACCAAGTCTCTGCTACCGTATAGATG	126	0.2153183635804368	No Hit
GGTTGGTTAGGGGTTTTTATGGGTTTGGAGATGGTGTTGTTAGACCAAGT	125	0.21360948767900476	No Hit
GTTTTATGGTGGGTTTTGGGAGTGGTTTAGACCAAGTCTCTGCTACCGTA	124	0.21190061177757272	No Hit
GGTGTGTTTGAGGTTGTAGGGATTAGGTGGGTTAGTTATTTTAGGAGATG	123	0.21019173587614068	No Hit
TGGTAGTTTAGGAGGTTTAGTTGGTGGAAATTTTGTGTTAGACCAAGTCTC	119	0.20335623227041252	No Hit
AGGTTTTGGGAAGTTTGGGTAGGGTAGGGAAGTTGGTAGACCAAGTCTCT	116	0.19822960456611644	No Hit
TGGGTGGTGGTTAGGGTTATGGATTTAAGAGAAGTGGTTAGACCAAGTCT	115	0.1965207286646844	No Hit
GGAATTTGAGGGTTGTTTTTAGATTTTTTCGGTTTTTTTTCGGTTGTTTA	111	0.18968522505895621	No Hit
TGGTAGTTTAGGAGGTTTAGTAGGGTGGGGGTTGGTGTGTAGACCAAGTC	111	0.18968522505895621	No Hit
AGAGGTTGGGTTTGGTTGGTAAGTTTTTTTTATATATAGAGAAGGGTTTT	108	0.1845585973546601	No Hit

TGTAGGGGATTATAGGGTAGGGTTTATGTGTTAAGTGTGTGTTATATGTG	106	0.18114084555179602	No Hit
GGAGGGGTAGGGTAAGGGTTTGGGTTTTTTTATGTTGTATGGTGTTTGT	105	0.17943196965036398	No Hit
TTGGGGATTTAGGGAGTTAGTTTTTAGTTTTGGTTTATGAATTTGTTTTA	104	0.17772309374893194	No Hit
AGGTTTTAGGTTTTGGGTGGGAATAGAGTATTAGACCAAGTCTCTGCTAC	103	0.17601421784749993	No Hit
ATGGGGGTGGAGGGGAATGGGGAGTTTAGGGAAGGTTTTTAGAGTTGGGT	103	0.17601421784749993	No Hit
TGGGAGAAGGGAAGATTGTGGTGAGGGTTTTGGAGTTAGACCAAGTCTCT	103	0.17601421784749993	No Hit
TGGTTTAGGGGAGAGGGGATTTTGCGTCGTGTTTTATTAAATTAAGGTGG	101	0.17259644604463583	No Hit
AAGGAGAATATGTGGGTAAGGTTTTATGGAGGATTTAGTTTGGGTAGGGT	101	0.17259644604463583	No Hit
TGGAAGTTTAATTTTGTTTAGAGTGGTGTGGTTTTGATGAATGGGGTAGA	98	0.1674698383403397	No Hit
TGTGTGTTTAGAGGGTGTAGGGCGGCGCGTAGAGGTTAGGGATGGTTTTG	98	0.1674698383403397	No Hit
AAGATGGGAAAGGTTTGGTGTTTATGTGGTGTTTGTTAGTTTTATAAGGG	97	0.1657609624389077	No Hit
GTTTTATGGTGGGGTGGGGAGTGGTTTAGACCAAGTCTCTGCTACCGTAT	97	0.1657609624389077	No Hit
GGAAGAGGAGGTGGTTGTTTTGGGGTAGACCAAGTCTCTGCTACCGTATA	96	0.16405208653747566	No Hit
GGGGATTAATAGTTGAAATGGTGAGGTAGTGGTGTGTATTGTGATTTGTT	95	0.1623432106360436	No Hit
AGGGAAGTTTGATTTTGTTTGGTTAGGTTTTTGGTTTGTGGTTAGACCAA	95	0.1623432106360436	No Hit
TGGTTAGGGGATGGTGGGAAGGAATTAGAATGGTATTCGAGTTTGCGTCG	94	0.16063433473461158	No Hit
AGTGAAGAGAGGAGGTTATTGAGATTACGATTATCGTGTTTTGGGTGATA	94	0.16063433473461158	No Hit
TTGGGGATTTAGGGAGTTAGTTTTTAGTTTTGGTTTACGAATTTGTTTTA	93	0.15892545883317954	No Hit
GAGAGTTGTTGTGGTTTTGGGATTGATTCGAGAAATGGAGGTTGTGATTT	92	0.15721658293174748	No Hit
GTTGGGAGTTGGGGGAAGAGGTAGGTTTAATTTTTTTTTTTTGGGTTTTT	91	0.15550770703031547	No Hit
GGTTGGTTAGGGGTTTTTATGGGTTGTAGTTGGAAAGAGGGGTTGGGTTT	91	0.15550770703031547	No Hit
AGGAAGTGGGGATTTTAAGTGGTTAAATGATTAGTTTAAGGTTACGTAGT	90	0.15379883112888343	No Hit
AGGTTGTAAGGGGTTTTGAGGAAATTAGTTATTTGATAGGTTTAGAGATT	90	0.15379883112888343	No Hit
TGGGGAGGGGGTAGAAAGAGTGTGGGTGTGGGGATAGATAGACCAAGTCT	90	0.15379883112888343	No Hit
GGGGATAGGAATGAGGGTGGGTTATAGGAGGTTTTAACGGGGAGGTAAAT	90	0.15379883112888343	No Hit
GGAGGTTGGGGAGAGTTTGGATTTGGAGGTGTGGGTAGACCAAGTCTCTG	88	0.15038107932601935	No Hit
ATGGGAGTGGGGGTTATGGTGGTTGTGGGAGTAGACCAAGTCTCTGCTAC	88	0.15038107932601935	No Hit
AGGGAGAGGGTTTTGGGGATAGTTTTTTTTGGGAATTTATATTGTGTTTTT	86	0.14696332752315527	No Hit
GGAGGTTGGGGAGAGTTTGGTTTAATTTTGTTTATCGTTATTTGGAGTTT	86	0.14696332752315527	No Hit
TGGTTGTTTAGGAAGGTTTAGGGTTTTTAGAATTGAGTGAGGGATAGACC	86	0.14696332752315527	No Hit
GGGGTTGTAGGGAGAAGGGAAGGTTTTGGTGGTTGGGAGGTATTTAGGGA	85	0.14525445162172324	No Hit
TGAGGTTGTGTTGAGTGAGAGATGTTTTTTTTTTAGTTTTTTGTAGATTT	85	0.14525445162172324	No Hit
TGTATAGGGTTATGGGGGTTGTGGGAGTAGACCAAGTCTCTGCTACCGTA	85	0.14525445162172324	No Hit
TAGTGGAGTGGGGGAGGTAGTGAGGGGAAGTTGAGGTAAATTTGGTGGTG	84	0.1435455575202912	No Hit
TGGGGATGGGAGTTTGGTTTAGTTGGTGTAAGTGTGGTTGGTTAGACCAA	84	0.1435455575202912	No Hit
TGGGTTTGGGATTTGGGTGTTGTTTTGGTTATATGATTATTGTTTGGTTT	83	0.14183669981885916	No Hit
TTGGGGATTTAGGGAGTTAGTTTATTATAGATTAGGTGTTTTTATATTTG	83	0.14183669981885916	No Hit
TGTTTGGGTTGTTTATTGGTGGTTTCGTATGAAGTTTCGGCGTTAGGTTT	83	0.14183669981885916	No Hit
GTGTTTGATTTTGGTTGGGGGAAGGTTAGACCAAGTCTCTGCTACCGTAT	82	0.14012782391742712	No Hit
TTGTAAGATGTGGGGATTGTGTGGAGGATGGAGAGGGATAGATGGTAGAC	82	0.14012782391742712	No Hit
GGTGGTAGTTTGGGAGTAGGAGTTGGAGATAGACCAAGTCTCTGCTACCG	80	0.13671007211456304	No Hit
AGTGTGAGAGTTTGGATGAGGTTGTAGTTGGAAGATGGGTTAGACCAAGT	79	0.135001196213131	No Hit
GTTGGGAGTTGGGGGAAGAGGTAGGTTTAATTTTTTTTTTTTTGGGTTTT	79	0.135001196213131	No Hit
TGTGTGTTTAGAGGGTGTAGGGAAGGTGTGGGAAGGATTAGACCAAGTCT	79	0.135001196213131	No Hit
TGGGGAAAGTGGAGATATAGTAGGAGGTAGTGTTTATGTGAGTTATGTTT	78	0.13329232031169896	No Hit
GGAAGAGGAGGTATAGATTTTGGGTAGACCAAGTCTCTGCTACCGTATAG	77	0.13158344441026693	No Hit
TGTAGGGGATTATAGGGTAGGGTTTATGTGTTAAGTGTGCGTTATACGTG	77	0.13158344441026693	No Hit
TGGGGAGGAGGGAGATATGGAAATAAAGAGGGTTAAGGTAAGAGGATGGA	76	0.1298745685088349	No Hit
TGGTTTAGGGGAGAGGGGATTTTGCGTCGTGTTTTATTGTAAATTAAGGT	74	0.1264568167059708	No Hit
TGGGTGGTTTTTAATGGGTTGTGTATTTTTTGAGATTTTGAGTGGAGGTA	74	0.1264568167059708	No Hit
AGAGTGTGGGTAGGGGTAGTTGTAGTTTTTTTGAGTTTTTTATTTTTTGT	74	0.1264568167059708	No Hit
TGTGGTAGAGTTGGTGGGAGTGTTTGTGAGTGTGTGGAGTGTGTATGTGT	73	0.12474794080453877	No Hit
TGGTAGAGTTGGGTTAGAATTTGGAGTGGTGTTGGGGTTGAGTAGTTAGA	72	0.12303906490310675	No Hit
TGGGTGGTGGTTAGGGTGATGGATTTAAGAGAAGTGGTTAGACCAAGTCT	72	0.12303906490310675	No Hit
GGGGATAGGAATGAGGGTGGGTTATAGGAGGTTTTAATGGGGAGGTAAAT	70	0.11962131310024265	No Hit
GAGAGTTGTTGTGGTTTTGGGAGGAAATTTGTTGTGGGTGATTATGGTAG	70	0.11962131310024265	No Hit
AGGTGGGGTTTGTTATTAGGGTTTGGGGAAGGTAGGATGTTAGACCAAGT	70	0.11962131310024265	No Hit
TGAGTTAGGGGGTAGGGTGGGGTGTAGTTTAGTTAATAGTATGTTAGGGC	70	0.11962131310024265	No Hit
AGTGAAGAGAGGAGGTTATTGAGATTACGATTATTGCGTTTTGGGTGATA	68	0.11620356129737859	No Hit
TAGTGGAGTGGGGGAGGTAGACCAAGTCTCTGCTACCGTATAGATGTCTC	68	0.11620356129737859	No Hit
TGTAGGGGATTATAGGGTAGGGTTTACGTGTTAAGTGTGTGTTATACGTG	68	0.11620356129737859	No Hit
TGTTTGGAAGAGGAGGAAATGAGGGTTATGGATTTAAGAGAAGTGGTTAG	67	0.11449468539594654	No Hit
TGTGTTTTGTAGAAAGGGTTTGGTTCGGAGGGTATTAGGAAAGGTTTAGT	67	0.11449468539594654	No Hit
TGTAGGGGATTATAGGGTAGGGTTTACGTGTTAAGTGTGCGTTATATGTG	67	0.11449468539594654	No Hit
GAGGGGGTTGATTGATGGGGAAATGAAGGTTATATGTTGTTAGACCAAGT	66	0.11278580949451451	No Hit
GGAATTTGAGGGTTGTTTTTAGATTTTTTTGGTTTTTTTTTGGTTGTTTA	66	0.11278580949451451	No Hit
TGGGAGAAGGGAAGATTGTGGGGGGTAGAATGTTTAGAATAGACCAAGTC	66	0.11278580949451451	No Hit
AAGAAAATGAAAATTGAGGTGGTGAAAAGGAATGAATGGGTTAGACCAAG	64	0.10936805769165042	No Hit
TGTGTGTGGTGGAGGTGTTTGTTTGTTAGTAGGGAAGTAGGGAAGAGAGT	64	0.10936805769165042	No Hit
ATATGGGTGTTGGGTGTGGTGAAAAGGAATGAATGGGTTAGACCAAGTCT	63	0.1076591817902184	No Hit
GGAAGAGGAGGTATAGATTTTGGGTAGGTTTAGACCAAGTCTCTGCTACC	63	0.1076591817902184	No Hit
AAGATGGGAAAGGTTTGGTGTGGGTAGTTTAGAGGTTTAGGAGTTAGACC	62	0.10595030588878636	No Hit
GTTGGGAGTTGGGGGAAGAGGGAGGGAATTTAGTGGTTGGGTTAATAGG	62	0.10595030588878636	No Hit
TGGGAGGTAGAGGTTGAGGTAGCGGTGGTTGGGATGGTTAGGAGAAAAGG	61	0.10424142998735433	No Hit
AGAGGGAGTGTTTTAAGTGGGTTGAAATATATTGTGAGTTAGACCAAGTC	60	0.10253255408592228	No Hit
TGGGTTGGGTTTGTTAAGTTGATTAGTGTGAATTGGTCGATTGAGGAAGG	60	0.10253255408592228	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	60	0.10253255408592228	No Hit
GGGAGGAATGAGGGAATGGGGAGGTAGTTTTGGTGGGTAGACCAAGTCTC	59	0.10082367818449024	No Hit

Adapter Content



Kmer Content



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
TGTTCTT	10	0.0070739286	145.0	145
GTGAAGA	115	0.0	126.08696	2
TATTGGA	60	0.0	120.833336	4
TGAAGAG	120	0.0	120.833336	3
GAATAGG	135	0.0	107.4074	9
AAGATGT	35	2.7291026E-7	103.57143	5
GAATATG	35	2.7291026E-7	103.57143	6
GAAAGAT	75	0.0	96.66667	8
TGTGAAA	30	1.5868418E-5	96.666664	3
GTATAGG	60	7.2759576E-12	96.666664	2
ATTGGAG	80	0.0	90.625	5
GGTTAAA	80	0.0	90.625	9
GTAAGAT	40	6.0453203E-7	90.625	3
AAGGATG	80	0.0	90.625	2
AAAGATG	80	0.0	90.625	9
ATATGGG	50	2.2313543E-8	87.0	1
TCGCGAG	60	8.112693E-10	84.583336	145
AGTAGAG	155	0.0	84.19355	8
GAAAATT	105	0.0	82.85714	9
GTGAAAG	35	3.41024E-5	82.85714	4

Produced by FastQC (version 0.11.5)

The End